Title
Visualization, Prediction, and Causal Inference: Applications in Healthcare

Author
Barter, Rebecca Louise

Publication Date
2019

Peer reviewed|Thesis/dissertation

Visualization, Prediction, and Causal Inference: Applications in Healthcare

by

Rebecca L. Barter

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Statistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Bin Yu, Chair
Professor Jasjeet Sekhon
Assistant Professor Avi Feller

Fall 2019

Visualization, Prediction, and Causal Inference: Applications in Healthcare

Abstract

Visualization, Prediction, and Causal Inference: Applications in Healthcare

by

Rebecca L. Barter

Doctor of Philosophy in Statistics

University of California, Berkeley

Professor Bin Yu, Chair

The recent wave of data collection in the field of healthcare has opened up an ocean of possibilities to learn and develop new exploratory, diagnostic, and prognostic methods. This thesis explores how three fields of statistics (1) data visualization, (2) prediction, (3) and causal inference, can help us leverage this data in order to answer a wide range of questions in healthcare.

Part I of this thesis presents a software package called *superheat* that can be used by researchers to visualize complex datasets and multi-faceted modeling results. The primary users of this software so far have been those in the medical research industry. In this thesis, we apply superheat in three case studies including (1) using a publicly available global organ donation database curated by the World Health Organization to understand and summarize the global organ donation trends, (2) visualizing groups of topics that appear in text data scraped from Google News, (3) examining model performance for a model designed to predict the brain's response to images using fMRI data. The theme of Part 1 of this thesis is visualization in healthcare.

Part II of this thesis introduces an analysis for predicting a patient's risk of developing a Surgical Site Infection (SSI) following surgery. A SSI is an infection that occurs at the site of a surgery within 30 days post surgery, and is responsible for up to 30% of hospital acquired infections. This method was developed in collaboration with healthcare professionals including infectious disease experts and surgeons at UC Davis. The theme of Part 2 of this thesis is prediction in healthcare.

Part III of this thesis presents a novel application of instrumental variables in causal inference, asking about the possible effectiveness of a "survival-benefit"-based liver transplant allocation scheme. The conclusion is that while there could be substantial benefit yielded from rethinking how organs are allocated, the feasibility of implementing such a scheme that relies drawing causal inferences from complex observational data is extremely difficult. The theme of Part 3 of this thesis is causal inference in healthcare.

To my family and friends.

Contents

List of Figures

List of Tables

Acknowledgments

This dissertation and my PhD would not have been possible without a great many people, the most important of whom is my advisor and mentor, Bin Yu. You believed in me and saw in me what others did not. Your patience, guidance, and care has helped me grow through these past few years and come out a stronger, wiser person with a clear purpose and direction. That is no small feat. You have truly been a role model, and I cannot thank you enough for the time and effort you put into my education and my life.

I would next like to thank my loving parents, Kerry and Philip, who always made me feel like I could *achieve* whatever I wanted to, while simultaneously allowing me to find my own path and *do* whatever I wanted to. You are both an eternal inspiration to me, and I feel lucky beyond words to be your daughter. I also want to thank my wonderful brother, Daniel, who led me down the road of mathematics in the first place, and without whom I probably would never have ended up studying statistics at Berkeley.

I am deeply thankful to Jas Sekhon for your advising and guidance, and to and Prabhu Shankar for being a wonderful collaborator throughout the SSI prediction project that forms Part II of this thesis. I'm also grateful to Karl Kumbier, Yotam Shem-Tov, and the entire Yu group for helpful discussions and friendship. Thank you Avi Feller and Peng Ding for sitting on my qualifying exam committee, and Avi for taking the time to read my dissertation. I'd also like to thank Jean Yang, whose support during my final years at The University of Sydney led me to Berkeley.

To all of my friends, especially Kellie Ottoboni, Wren Suess, Mike Chapman, JB Chapman, Chris Lalau Keraly, Mika Endo, and many, many more - thank you all for supporting me, and teaching me how to live life to the fullest. You have all made these past few years the best years of my life.

A big thank you to all of the folks at BIDS who made our workspace feel like a home and opened my mind to different perspectives and the never-ending discussion of what data science is and how to do it well. My PhD experience was greatly enriched thanks to my time as a BIDS fellow. Thanks also to CITRIS for providing the grant that supported the Surgical Site Infections project that forms Part II of this thesis.

Finally, we are so lucky to have such wonderful staff members in the Statistics department, especially La Shana, Mary, Laura and Keyla, who somehow make sure everything keeps running smoothly.

My time at Berkeley have been the best years of my life, and I will carry them with me always.

Chapter 1

Introduction

The quantities of data collected by the healthcare industry are vast. These data present multi-dimensional views of hospitals, patients, and diseases, and come from a range of sources including

- Electronic Health Records (EHRs) collected by hospitals on their patients containing laboratory information, diagnosis information, medication information, diagnosis information, and more.
- Clinical trial data, often 'omics datasets on patients' genomes, proteomes, microbiomes, etc collected from trial participants to answer specific questions designed to further our knowledge of human health and medical efficacy.
- Administrative and Claims data collected by hospitals and insurance companies.
- Wearable technology data collected on people who wear devices, such as smartwatches and heart rate monitors, that collect information about movement, heart rate, and more.

While there are already volumes of healthcare data from each of these sources and more being produced every day, the rate at which we, the researchers, are able to learn from the data is lagging. One of the principal barriers stunting our ability to learn from this ocean of healthcare data is privacy. It is critically important that the privacy and anonymity of the patients on which the data is collected is preserved and respected. Without healthcare data privacy and protection, patients with chronic conditions may unfairly face social, professional, and financial disadvantage.

As such, healthcare data is rarely publicly available, and when it is, it is often provided without the annotating labels required to answer most questions. As researchers, the only way to access these vital data sources is to work directly with those who collected the data, or to get express permission from the organization that collected the data. This can be a tedious and tiresome process. The data in this thesis came from many different sources including:

- Publicly available organ donation data from a World Health Organization (WHO) database aggregated by country (and so does not contain any personal information).
- Individual-patient level data collected in UC Davis' EHR database, and additional hospital databases specifically relating to surgical information. This data was obtained via our collaborators in the School of Medicine at UC Davis.
- Individual-patient level data made available by request from the United Network for Organ Sharing (UNOS); the organization that governs organ donation in the US.

While obtaining the public data was easy (since it was not individual-patient level data), obtaining the other two sources of data were substantially more difficult, and in each case took close to a year. The difficulties that researchers face when trying to access healthcare data, while being realistic in terms of preserving patient privacy unfortunately means that the rate of learning is slow, and so thus is our ability to drive the technology future of healthcare.

However, even without data access issues, researchers developing methodology and technology for use by healthcare professionals face additional hurdles. Simply developing algorithms based on data is not enough. Many researchers incorrectly assume that if they make a nice visualization tool or predictive method it will magically be adopted by the healthcare industry. However, much of the time, the software is not open source or directly usable, the models are not properly validated on sufficiently diverse populations, no graphical user interface (GUI) is produced, and the algorithms produced aren't directly useful to clinicians in the first place. Algorithms and technologies need to be developed in collaboration with the healthcare professionals who would be using it. Feasible means of implementation that does not add to the burden of the healthcare workers need to be prioritized.

Beyond ensuring that the algorithms and technology developed by researchers are useful, they also needs to be trustworthy. The stakes are higher in healthcare than in the tech industry: the stakes are peoples lives. It is critical that models be built on substantially varied cohorts of patients that represent real populations, and are validated widely across a variety of different scenarios. However, the reality is that most researchers do not have the resources nor the bandwidth to do this.

The most traditional strains of research that can enact real change in the healthcare industry are clinical trials. However, even these can span decades and are not problem-free. Clinical trials are expensive, often based on populations that are not representative of the general population, and face complex and burdensome regulations. While clinical trials will continue to be play a critical role in the development of new medical knowledge, we are at a moment in time where we have the data and computational power to make groundbreaking findings, and develop game-changing technologies, we just need to figure out how to use it safely, effectively, and ethically.

This thesis consists of three parts. In Part I, the theme is **visualization in healthcare**, and we introduce our ready-to-use open-source visualization software called *superheat* that is already being used by thousands of people in their research. In Part II, the theme is

prediction in healthcare, and we collaborate with surgeons and infections disease experts to develop algorithms for predicting infections arising at the site of a surgery within 30 days, known as a *Surgical Site Infection (SSI)*. In Part III, the theme is **causal inference in healthcare**, and we draw causal inferences about a benefit-based liver transplant allocation system that has the potential to dramatically impact the organ transplant system.

The remaining sections in this introductory chapter provide a broad overview of how visualization, prediction, and causal inference are currently being used in healthcare, the challenges we face, and where we might go from here.

1.1 Visualization in healthcare

Data visualization plays a pivotal role in illuminating the underlying structures present in data, as well as interpreting the process and performance of data-driven models. As humans, we do not speak the same language as computers. We can, however, use visualization to translate the computer's representation of data and models into a visual language that we can comprehend. Visualization thus provides one of the core components of human-computer interaction, allowing us to capture and understand complex patterns that would be otherwise impossible for us to mentally digest.

As the datasets being produced and the models being built in the field of healthcare become more and more complex, our ability to visualize and thus understand them has diminished dramatically. Simple low-dimensional scatterplots, bar plots, histograms, and boxplots are woefully inadequate for visualizing information that lives across hundreds or thousands of dimensions.

Many insights are often missed because the analyst didn't take time to properly visualize their data. Data visualization can be used both for the analyst to understand the data (exploratory data analysis), as well as to communicate results and findings to an external audience (explanatory data analysis) [74].

The primary existing methods for visualizing multi-dimensional datasets are heatmaps, parallel coordinate plots [41], and multi-panel plots such as scatterplot matrices [17]. However, parallel coordinates and multi-panel plots fail to present more than 20 or so dimensions at once, whereas heatmaps can be extended to thousands of dimensions. A heatmap can be used to visualize a data matrix by representing each matrix entry by a color corresponding to its magnitude, enabling the user to visually process large datasets with thousands of rows and/or columns.

In Part I of this thesis, we present our R package, superheat [8], for visualizing complex datasets using flexibly extendable heatmaps that use intuitive color transitions, and combine the heatmap with scatterplots, bar plots, line plots and more. The greater ease of implementation, flexibility of customization, and visual attractiveness of *superheat* sets our software apart from its static competitors.

While only the first part of this thesis focuses specifically on data visualization, visualization in general plays a significant role throughout all three parts of this thesis.

1.2 Prediction in healthcare

If clinicians could adequately harness the information contained in electronic health records (EHR) and other data sources, their treatment decisions could be driven not only by their own professional experience and judgement calls, but also by the outcomes experienced by other similar patients subject to each treatments. However, in order to develop a clinical decision support (CDS) tool that can be used by clinicians, there is a long road that must be travelled. Many data-driven problems in healthcare can be formulated as prediction problems, wherein the patterns captured by the data collected from past patients be pooled together to predict responses for new patients. Responses of interest might be survival time, recovery rate, quality of life improvement, length of hospital stay, or a myriad of other things.

Unfortunately, developing predictive methods is not quite as easy as throwing an entire database into an arbitrary predictive algorithm. There are many things that must be done first. For instance, the outcome of interest needs to be formulated thoughtfully and clearly. If your goal is to reduce the rate of hospital acquired infections (HAIs), you might want to tackle this by developing an approach for identifying patients at risk of HAI so that they can be more closely monitored. However, HAIs come in many forms, and it might make sense to focus on just one type, such as infections arising from surgery, known as Surgical Site Infections (SSI). If you're focusing on SSIs then, you also need to specify the time period during which you define the infection: for instance, the infection must arise within 30 days following the surgery for it to be considered an SSI.

Next, you need to decide what data you want to use. This may be dictated by the data that you can access, but even if you have abundant access to data, you will need to identify what data is *relevant* to your question. For instance, will you need all data from the month before surgery, or perhaps you need to consider the past 6 months, or even the past year? Sometimes technical components will dictate your decision, such as, how far back do you need to go so that each patient has at least one data point? Other times, domain knowledge will play a strong role: it might be very unlikely that lab measurements taken a year ago will bear any relationship to an infection arising from a surgery today. At the end of the day, many judgement calls will need to be made, but these judgement calls will typically be informed by the data, domain knowledge (either by the analyst or domain expert collaborators), and experience. Even once you've collected your data, and have gone through the iterative process of moulding and cleaning it, new data may be collected, or previous decisions, or even the question itself, might be refined.

Only after implementing the process of problem formulation, data collection, and data cleaning is it time to implement a predictive algorithm. Usually this will come in the form of a Machine Learning (ML) algorithm that can use the patterns present in the current data to predict the response for future data. However, just building a ML model isn't enough. There are a many big challenges when it comes to implementing ML models in practice. Differences between the population of patients that were used to build the model and the population that it is later applied to can result misleading predictions that can have dramatic consequences. For example, several genetic studies have been criticized for not accounting

for genetic diversity in non-European populations, leading to misdiagnoses in patients with African and unspecified ancestry [66, 32]. This is an excellent example of where the role of humans in ML is critical. Humans need to step in to ensure that our ML algorithms are behaving in a fair, ethical and unbiased way.

Another challenge facing ML is sample size. In order for predictive models to be able to accurately capture the patterns that relate the predictive features to the response, there needs to be enough data for the patterns to be distinguishable from noise [51]. This is especially relevant in imbalanced datasets or rare-event prediction problems, where even if the overall sample size might seem large, the class you're trying to predict may not contribute many samples to the data [32].

A more recent challenge is the reputation of ML algorithms as an uninterpretable "black box". Why should medical practitioners trust an algorithm that doesn't explain how or why it made its predictions? This is particularly true of Deep Learning algorithms that have seen wide success in a range of applications[61, 78], but are notoriously difficult to interpret. While there have been some recent efforts to interpreting Deep Learning algorithms applied to images and text [72], we are still a long way from these models being interpretable in general applications.

Finally, implementing ML algorithms involves more than running a model on the researcher's computer. The final stage of the pipeline which involves generating a Graphical User Interface (GUI) is often overlooked, and as a result, the vast majority of predictive models that apparently work so well are never implemented in practice [93].

In Part II of this thesis, we develop a predictive model for predicting surgical site infections based on the complete cohort of patients that underwent surgery at UC Davis from 2014 through to the end of 2017. This project was conducted in direct collaboration with medical informaticians, infectious disease specialists, and surgeons at the UC Davis Medical School. While we have only so far developed a predictive model that has been internally validated, we will soon be conducting external validation on a separate set of patients, as well as developing a GUI so that the surgeons at UC Davis can use our model's predictions to drive their medical decisions, as well as provide feedback on its performance.

1.3 Causal Inference in healthcare

Causal inference as a field of study is all about showing that interventions *cause* a response, rather than just being *associated* with a response. These two concepts are frequently confused for one another.

In Part II of this thesis, we find that the length of surgery is a strong predictor of surgical site infection. Does that mean that the increased length of surgery *causes* infections? Not necessarily, although many would quickly jump to that conclusion, especially because it is easy to hypothesize *why* the increased length of surgery could lead to infection e.g. there is more time - and thus opportunity - for bacteria to enter into the patient's open wound. However, as a potential counter-point, what if patients with compromised immune systems

also tended to have longer surgeries? Then it might actually be the lack of a strong immune system in the patient, rather than the length of the surgery that led to the infection. Unless you can rule out *every single alternative possible cause* of the increased rate of infection among patients with longer surgeries, you cannot conclusively say that it was indeed the length of the surgery that caused the increased risk of infection.

The only way it is possible to truly conclude that the longer surgery caused an increased risk of an infection is to compare the rate of infection among a group of patients if they were to undergo two identical surgeries, but where one surgery was half an hour shorter than the other. However, since you cannot perform two identical surgeries at once on the same patient, this is unfortunately impossible. That it is impossible to perform and observe the outcome of these two possible surgeries at once forms the *fundamental problem of causal inference*.

The most accepted way to get at the causal effect of an intervention (e.g. the effect of length of surgery) on an outcome (e.g. infection) is instead to compare the outcome in two equivalent groups of patients: one group that has shorter surgeries and one group that has longer surgeries. If indeed everything else is the same between these two populations (e.g. both groups of people were equally sick, had the same surgical conditions, and equivalent surgeons), then any difference in infection rate that is observed *must* be due to the different surgery lengths. This idea forms the basis of clinical trials in which a group of people are *randomly* split into two different groups: one that receives intervention A (e.g. shorter surgeries) and the other that receives intervention B (e.g. longer surgeries). The randomness of the group allocation is what creates the equivalence of the two intervention groups.

However, when you are unable to implement a random intervention (i.e. manually decide the length of the surgery), as is usually the case, you can try to infer causation from observational data (e.g. electronic medical records that record infection rates and surgery length). The problem with observational data is that the intervention (surgery length) was almost certainly not random, and so there are probably characteristics of the patient, surgeon, hospital, etc that contribute both to the length of surgery and the risk of infection. Such characteristics are called *confounders*. If, however, you can identify all of the confounders that influence both the length of surgery *and* the risk of infection, then you might be able to argue that conditional on those features, the length of surgery is random, and you can proceed as in the random experiment case as above. This is only possible, however, if you both know what these confounders are, *and* they are measured in your data.

Part III of this thesis solves a real problem that involves inferring causation when there are *unmeasured* confounders present. Specifically, we will use a method called instrumental variables to infer whether receiving a liver transplant earlier leads to increased survival post-transplantation.

Part I

Visualization: The Superheat R Package for Visualizing Complex Data

Chapter 2

The superheat R package

2.1 Introduction

The rapid technological advancements of the past few decades have enabled us to collect vast amounts of data with the goal of finding answers to increasingly complex questions both in science and beyond. Although visualization has the capacity to be a powerful tool in the information extraction process of large multivariate datasets, the majority of commonly used graphical exploratory techniques such as traditional scatterplots, boxplots, and histograms are embedded in spaces of 2 dimensions and rarely extend satisfactorily into higher dimensions. Basic extensions of these traditional techniques into 3 dimensions are not uncommon, but tend to be inadequately represented when compressed to a 2-dimensional format. Even graphical techniques designed for 2-dimensional visualization of multivariate data, such as the scatterplot matrix [21, 4] and parallel coordinates [42, 43, 44] become incomprehensible in the presence of too many data points or variables. These techniques suffer from a lack of scalability. Effective approaches to the visualization of high-dimensional data must subsequently satisfy a tradeoff between simplicity and complexity. A graph that is overly complex impedes comprehension, while a graph that is too simple conceals important information.

An existing visualization technique that is particularly well suited to the visualization of high-dimensional multivariate data is the heatmap. Today, heatmaps are widely used in areas such as bioinformatics (often to visualize large gene expression datasets, for example in [97] and [95]), yet are significantly underemployed in other domains. There exist a wide range of standard heatmap software available, including inbuilt R functions such as *image* and *heatmap*, as well as functions from R packages such as *heatmap.2* from the *gplots* package, *heatmap.3* from the *GMD* package, the *pheatmap* package [52] and its extension *aheatmap* [31] from the *NMF* package.

A heatmap can be used to visualize a data matrix by representing each matrix entry by a color corresponding to its magnitude, enabling the user to visually process large datasets with thousands of rows and/or columns. While the computational power of the 21st century

has enabled researchers to produce increasingly rich and complex heatmaps, the earliest sources of the heatmap date back to at least the 1800s, where [62] used color to represent the numerical values of various social statistics in Paris.

Even the modern practice of emphasizing structure in the data by clustering together similar rows and columns of the heatmap is not new. Authors such as [14] used such techniques over 100 years ago to highlight relationships in educational data. A more recent development is the practice of appending a dendrogram to the rows and/or columns of a heatmap to present the hierarchy of clusters in the data. Authors such as [59, 35] and [19] originally developed heatmaps that displayed both the reordered/clustered data matrix as well as adjacent diagonal similarity matrices with dendrograms attached. These more complex (but perhaps more informative) versions of the clustered heatmap later morphed into the more common version we see today which appends the dendrograms directly to the clustered data matrix [109, 25]. A more detailed history of the heatmap is provided in [110].

While augmentation by a cluster dendrogram has been fairly common practice for the past two decades, it remains fairly uncommon to augment heatmaps by other types of information. Recently, interest has arisen in combining heatmaps with other traditional plot types such as barplots, scatterplots, and histograms, and several authors have produced software for producing such visualizations such as the *ComplexHeatmap* [37]. Another recent avenue for expanding the traditional heatmap toolbox is the incorporation of hover and click interactivity such as in *heatmaply* [29]. Both interactivity and additional subplots have been combined in the *iheatmapr* R package by [89]. Our R package, *superheat* [8], was one of the early packages (originally developed in 2015) to incorporate additional information in the form of adjacent subplots such as barplots, boxplots, line plots, scatterplots and more. The greater ease of implementation, flexibility of customization, and visual attractiveness of *superheat*, as we will show throughout this Chapter, sets our software apart from its static competitors.

2.2 Choosing row/column ordering and color mapping in heatmaps

While heatmaps can be incredibly useful for visualizing large matrices, they can also be misinterpreted if designed improperly [22]. Two features of the heatmap most likely to lead to misrepresentation of the data are (1) the choice of row/column ordering in generating clustered heatmaps, and (2) the choice of color mapping. In this section, we will provide some brief advice on the use of for row/column ordering, and discuss how a quantile color mapping helps alleviate issues that can arise when manually the choosing the scale for a heatmap color map.

Row and column ordering

Interpretation of a heatmap can vary based on the ordering of the rows/columns, so it is always a good idea to ensure that any patterns highlighted by a clustered or re-ordered heatmap are stable. Our recommended approach to assess the stability of the patterns identified in such heatmaps is to re-generate the heatmap on various random subsets of the data to ensure that the patterns identified are consistent [114].

Color maps

A color map consists of two components: (1) the choice of color space/scheme, and (2) the functional mapping that dictates which color (within the specified color space) each data point is mapped to. Care must be taken when defining the color scheme and mapping [111].

It is important that the selected heatmap color space is perceptually uniform, i.e. the difference between two colors, as perceived by the human eye, is proportional to the Euclidean distance between the two colors in the color space. The default color scheme for our *superheat* package, *viridis*, is perceptually uniform. It has been shown, however, that many of the popular color schemes, such as the "rainbow" color scheme, are not [94, 15].

Having selected an appropriate color scheme, the next decision is how to map data into the corresponding color space. For the majority of heatmap software, the default functional mapping from data to color is linear: equal distances in data space are represented as equal distances in color space. In most datasets, however, the data is not spread uniformly throughout the range from the smallest to the largest value. Instead, the data might be more dense in the middle of the range, or be skewed towards larger or smaller values. In this case, a linear mapping from data space to color space will highlight outliers by emphasizing the data points that have largest distances from the bulk of the data. As a result, it is common in practice for users to manually adjust the color transition positions until they feel that they have highlighted as many trends in the center of the data as possible. Manual selection of the data-to-color mapping can potentially lead to a scenario in which the researcher is simply highlighting noise or is unintentionally hiding information. For example, a color map that represents all negative values as black and all positive values as ranging from dark blue to light blue will hide the information contained within the negative data space.

Figures 2.1 and 2.2 present two heatmaps, each using the same viridis color space but different color mappings: Figure 2.1 implements a linear color map, while Figure 2.2 implements a quantile color map (described below). The data underlying the heatmaps comes from lawyers' ratings of a subset of 20 state judges in the US Superior Court from the New Haven Register in 1977. These ratings were collected on 12 characteristics: contacts (the number of contacts of the lawyer with the judge), judicial integrity, physical ability, demeanor, diligence, case flow managing, prompt decisions, worthy of retention, preparation for trial, familiarity with law, sound oral rulings, and sound written rulings. The data can be found as a part of the inbuilt `datasets` package in R.

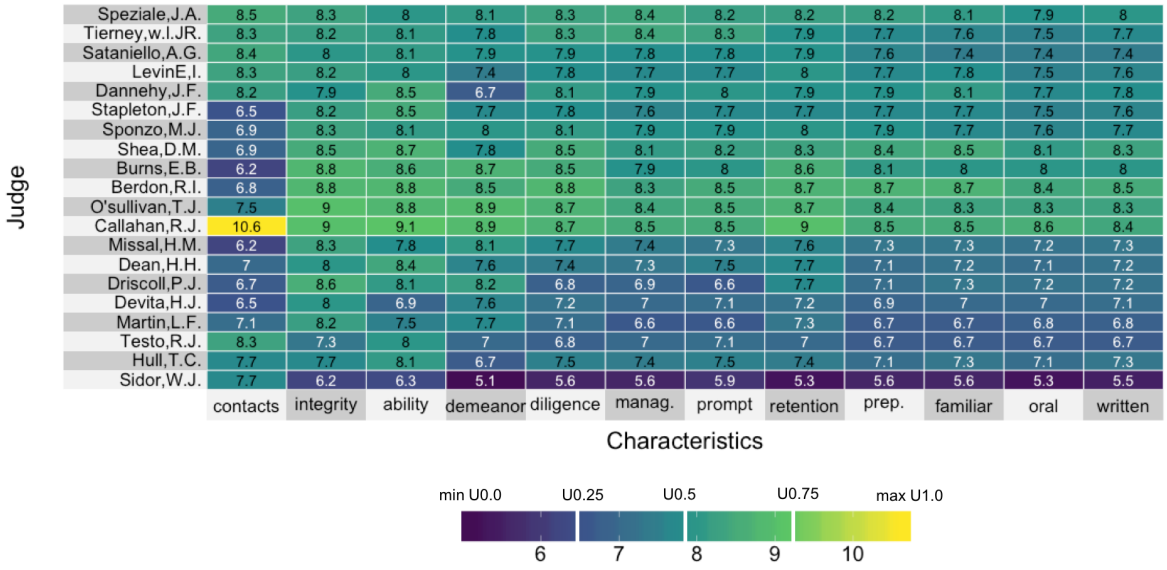

Figure 2.1: A heatmap with a viridis color space and linear color map of the lawyers' ratings of 20 state Judges in the US Superior Court. The white vertical bars in the legend represent the positions of three central (equidistant) colors in color space.

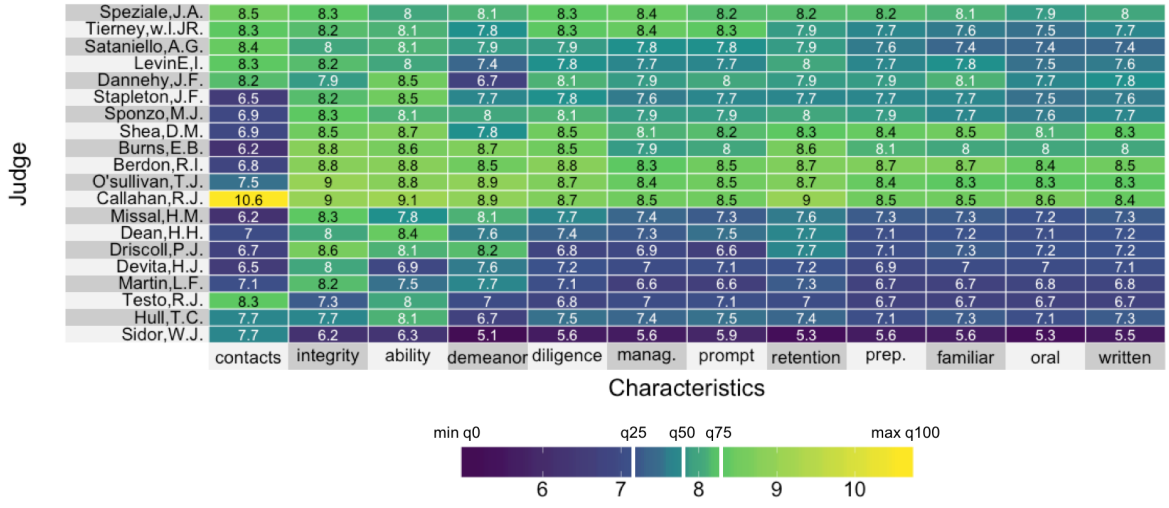

Figure 2.2: A heatmap with a viridis color space and quantile color map of the Lawyers' ratings of 20 state Judges in the US Superior Court. The numbers in the cells show the actual ratings. The white vertical bars in the legend represent the same three colors from Figure 2.1, but their positions are mapped from the 25th, 50th, and 75th data quantiles.

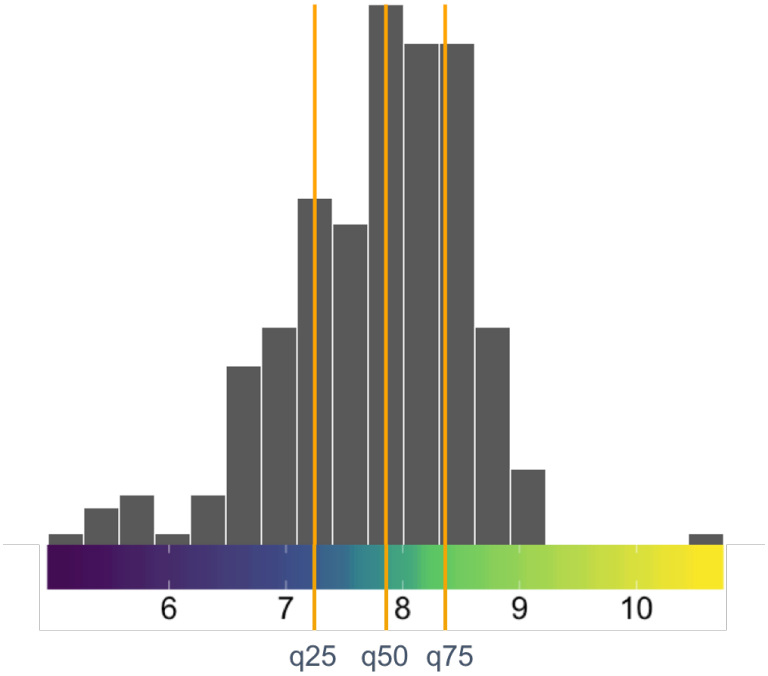

Figure 2.3: The distribution as a histogram of the lawyer's ratings on US superior court judges placed on top of the quantile color map (from Figure 2.2). The quantiles are highlighted by vertical orange lines.

With a linear mapping (Figure 2.1), color is distributed uniformly throughout the range of the data-to-color map (as represented by the equidistant vertical white bars in the legend representing three equidistant colors in color space). As a result of the linear mapping, there is a notable lack of contrast among the bulk of the data in Figure 2.1. Since the data are far from uniformly distributed (80% of the data lies between the values 7 and 9), most of the ratings are presented as being very close together, which, relative to the range of the data, they are. Unfortunately, this feature of the linear mapping makes it very difficult to tease out the patterns in the data when the majority of the data values do not uniformly span the range of the data. Linear color maps would be appropriate if the user wanted to highlight the outliers and subdue patterns within the region of typical data values, however, this is usually not the goal of a heatmap.

Figure 2.2 shows the same heatmap with an alternative quantile color mapping which allows for quicker transitions between the colors in regions where the bulk of the data lie. The quantile color map uses the quantiles of the data to dictate where the color transitions should take place within the heatmap. In a color space that is defined by a set of five sequential colors, the first color is centered at the minimum value in the data (or the 0th quantile), the second color is centered at the 25th quantile, the third color is centered at the median of the data (the 50th quantile), the fourth color is centered at the 75th quantile, and the fifth color is centered at the maximum value of the data (the 100th quantile). The

transitions from one color to the next happen in between these quantiles. The positions of three central colors that are equidistant in color space (but whose mapping is defined by the 25th, 50th and 75th quantiles in the data) are presented as white bars in the legend of Figure 2.2. Compare the positions of the same central colors in Figure 2.1.

Notice that are more distinct groupings visible in Figure 2.2 as compared to Figure 2.1. Quantile color maps are appropriate for users who want to highlight the typical data values and reduce the influence of outliers on the color transitions. Figure 2.3 shows the distribution of the data with the quantile color map.

While alternatives to heatmaps such as scatterplot matrices and parallel coordinate plots are less sensitive to choices of color and order, they quickly become intractable in the presence of even tens of variables. Heatmaps are able to display substantially more information in less space than either of these popular counterparts. A detailed comparison of the heatmap with scatterplot matrices and parallel coordinates will feature in our first case study in Section 2.4.

2.3 The superheat R package

Inspired by a desire to visualize a design matrix in a manner that is supervised by some response variable, we developed an R package *superheat* (short for "supervised heatmap") for producing "supervised" heatmaps that extend the traditional heatmap via the incorporation of additional information. Superheatmaps are flexible, customizable and very useful for presenting a global view of complex datasets. Such plots would be difficult and time-consuming to produce without the existence of software that can automatically generate the plots given the user's preferences. *Superheat*, builds upon the infrastructure provided by the *ggplot2* [106] R package to develop an intuitive heatmap function that possesses the aesthetics of *ggplot2* with the simple implementation of the inbuilt heatmap functions. While *ggplot2* itself contains functions for producing visually appealing heatmaps, it requires the user to convert the data matrix to a long-form data frame consisting of three columns: the row index, the column index, and the corresponding fill value. Although this data structure is intuitive for other types of plots, it can be somewhat cumbersome for producing heatmaps. For this reason, *superheat* accepts matrix inputs directly and does not make use of the *ggplot2* grammar of graphics [105].

Below we highlight some key features and usage of superheat. Readers looking for more details can find the wide variety of features as well as extensive instructions on usage of *superheat* in the online Vignette (see supplementary materials of [8] for the URL). The superheat package contains a single function: the self-named `superheat` function. The data matrix to be plotted is to be provided as the first argument, `X`. All other arguments of the `superheat` function are optional, and some are described below.

Adding additional information

One of the primary components of *superheat* is the ability to add additional sources of information in the form of scatterplots, barplots, boxplots, line plots, and dendrograms adjacent to the rows and columns of the heatmap. These adjacent plots allow the user to explore their data to greater depths, and to take advantage of the heterogeneity present in the data to inform analysis decisions. Some examples of the basic structure of a superheatmap are presented in Figure 2.4.

To add a plot above the heatmap, the user provides a vector to the `yt` ("y top") argument, where the length of the vector is equal to the number of columns in the heatmap. Similarly, to add a plot to the right of the heatmap, the user provides a vector to the `yr` ("y right") argument.

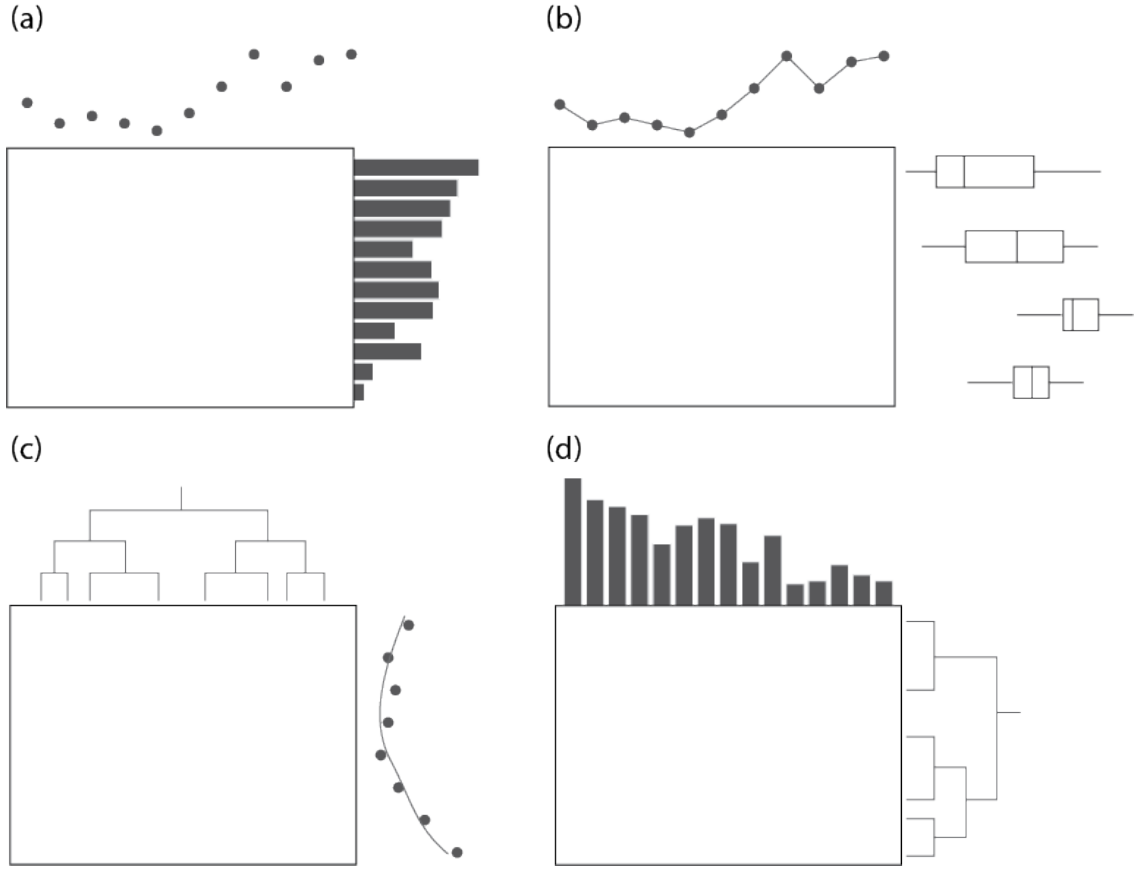

Figure 2.4: Four examples of *superheat* layouts. Panel (a) shows a scatterplot added to the columns, and a bar plot added to the rows. Panel (b) shows a scatter-line plot added to the columns and grouped boxplots added to the rows. Panel (c) shows a dendrogram added to the columns and a scatter-smooth plot added to the rows. Panel (d) shows a bar plot added to the columns and a dendrogram added to the rows.

The type of plot can be specified by setting the `yt.plot.type` or `yr.plot.type` argument to `'scatter'`, `'bar'`, `'boxplot'`, `'scattersmooth'`, `'smooth'`, `'scatterline'`, or `'line'`. Note that boxplots can only be added when the rows or columns are grouped (see Section 2.3). Overlaid text such as the data itself can be added to the heatmap using the `X.text` argument. Row or column dendrograms can be added by setting the `row.dendrogram` or `col.dendrogram` to be `TRUE`.

Specifying row/column ordering and grouping

By default, *superheat* does not reorder the rows or columns of the matrix provided. The order of the rows and columns (and simultaneously the data in the adjacent plots) can be changed by providing the `order.rows` and `order.cols` arguments with an index vector specifying the position of the columns/rows. For users that would like *superheat* to automatically rearrange the rows/columns in order to highlight structure, setting the arguments `pretty.order.rows` and `pretty.order.cols` to `TRUE` will apply a hierarchical clustering algorithm and rearrange the rows/columns accordingly.

Superheat has inbuilt clustering capabilities wherein the user can specify the number of row or column clusters they would like using the `n.clusters.rows` and `n.clusters.cols` arguments. *Superheat* will then run a k-means (the default clustering algorithm) on the data matrix and will group together the rows or columns that are in the same cluster (while respecting the order of the rows/columns specified by `order.rows` and `order.cols` within each cluster). To select the number of clusters, it is recommended that the user does so prior to the implementation of the superheatmaps using standard methods such as Silhouette plots [84]. Users can also provide their own cluster membership vectors using the `membership.rows` and `membership.cols` arguments.

When using clustering within *superheat*, the resulting heatmap is a "grouped" heatmap, to which boxplots and aggregate bar plots can be added as an adjacent plot for each group of rows or columns. Grouped heatmaps with a large number of rows/columns can be smoothed so that each row/column group is presented by a single color corresponding to the median entry, rather than to show each matrix entry individually. An example of a grouped heatmap with smoothing can be seen in our second case study in Figure 2.9.

Specifying the color map and color scheme

By default, *superheat* uses a quantile color map (see Section 2.2) with the perceptually uniform viridis color scheme. Users who wish to deviate from the default viridis quantile color map can specify their own color palette using the `heat.pal` argument, or users can choose alternative color schemes from among the sequential color brewer [39] schemes (setting `heat.col.scheme` to one of `'red'`, `'purple'`, `'blue'`, `'grey'`, `'green'`). Users can manually specify a data-to-color map using the `heat.pal.values` argument, which expects a vector whose length equals `heat.pal` and which specifies the center position of each color specified in `heat.pal`. For example, if we have a dataset whose minimum value is 0 and

whose maximum value is 10, if we set `heat.pal = c('white', 'blue', 'black')` and `heat.pal.values = c(0, 0.2, 1)`, then the data value of 0 will map to "white", the data value of 2 will map to "blue", and the value of 10 will map to "black", with linear transitions between each of these colors.

Further implementation information

The development page for *superheat* is hosted on GitHub, where the user can also find a detailed Vignette describing further information on the specific usage of superheat as well as a host of options for functional and aesthetic customizability. Details of the analytic pipeline and code for the case studies presented in this Chapter can be found in [8].

The remainder of this Chapter will present three case studies that highlight the ability of *superheat* to (1) combine multiple sources of data together, (2) uncover correlational structure in data, and (3) evaluate heterogeneity in the performance of data models.

2.4 Case study I: combining data sources to explore global organ transplantation trends

The worldwide demand for organ transplantation has drastically increased over the past decade, leading to a gross imbalance of supply and demand. In the United States, there are currently over 100,000 people waiting on the national transplant lists but there simply aren't enough donors to meet this demand [1]. This imbalance is worse in some countries than others as organ donation rates vary hugely from country to country, and it has been suggested that organ donation and transplantation rates are correlated with country development [30].

This case study will explore combining multiple sources of data in order to examine the recent trends in organ donation worldwide as well as the relationship between organ donation and the Human Development Index (HDI). The organ donation data was collected from the WHO-ONT Global Observatory on Donation and Transplantation, which represents the most comprehensive source to date of worldwide data concerning activities in organ donation and transplantation derived from official sources. The database (available from [33]) contains information from a questionnaire annually distributed to health authorities from the 194 Member States in the six World Health Organization (WHO) regions: Africa, The Americas, Eastern Mediterranean, Europe, South-East Asia and Western Pacific.

The HDI was created to emphasize that people and their capabilities (rather than economic growth) should be the ultimate criteria for assessing the development of a country. The HDI is calculated based on life expectancy, education and per capita indicators and is hosted by the United Nations Development Program's Human Development Reports (available from [99]).

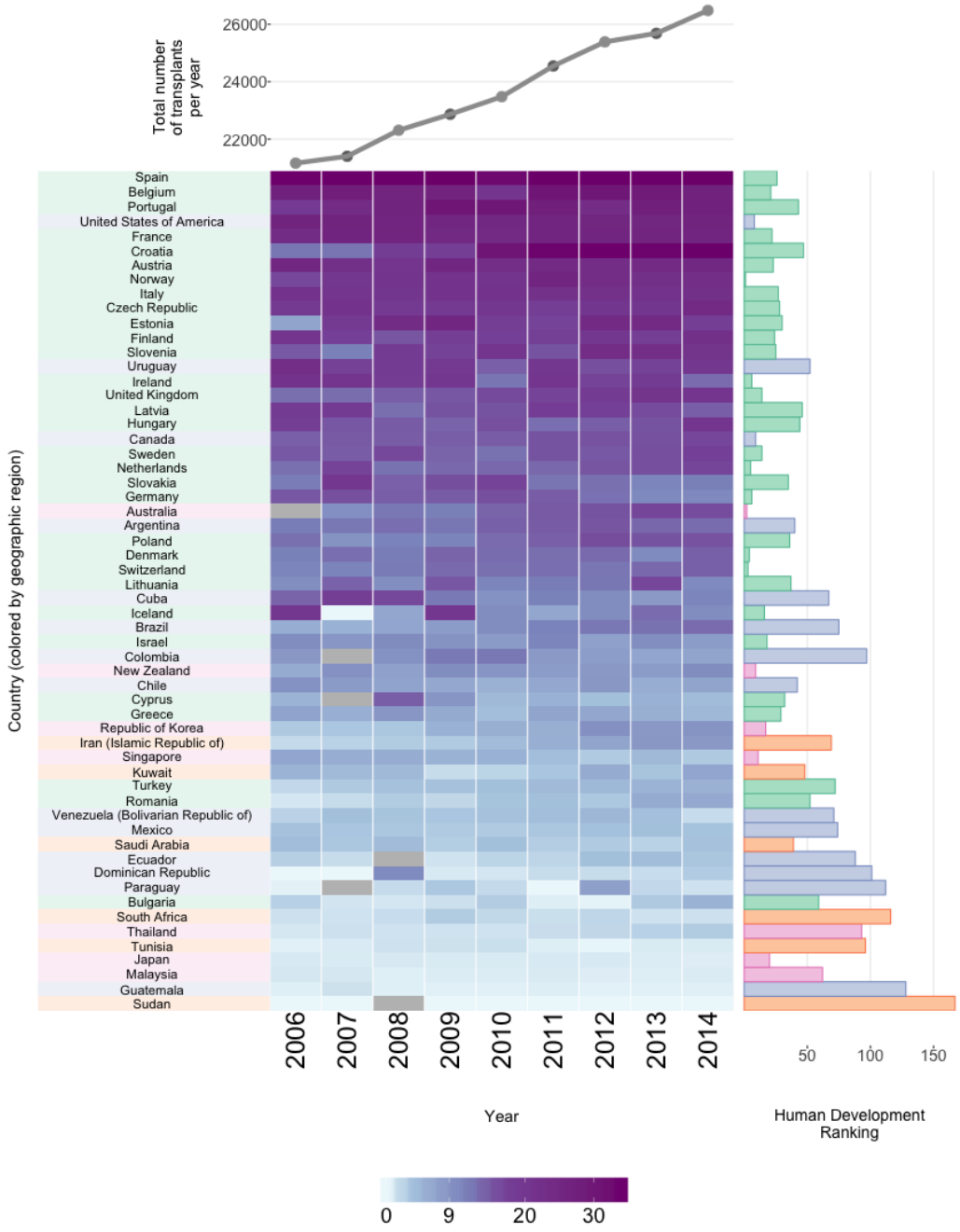

Figure 2.5: Organ donations and HDI by country. The right-hand bar plot displays the HDI ranking (lower is better). Each heatmap cell shows the number of organ donations from deceased donors per 100K. Grey cells correspond to missing values. The rows (countries) are ordered by average transplants per 100K. The country labels and HDI bar plot are colored based on region: Europe (green), Eastern Mediterranean (purple), Western Pacific (yellow), America (orange), South East Asia (pink) and Africa (light green). The upper line plot shows total organs donated per year.

Exploration

In the superheatmap presented in Figure 2.5, the central heatmap presents the total number of donated organs from deceased donors per 100,000 individuals between 2006 to 2014 for each country, restricting to countries for which data was collected for at least 8 of the 9 years.

Note that relaxing the country inclusion requirement to available data for 7 of the 9 years would include some additional countries (Bhutan, Costa Rica, Kenya, Luxembourg, Myanmar, Nigeria, Oman, Panama and the Syrian Arab Republic), however, in the interests of space, we do not include these. Further note that there are several countries (China and India included) for which there is no total deceased organ donor data available.

Above the heatmap, a line plot displays the overall number of donated organs over time, aggregated across all 58 countries represented in the figure. We see that overall, the organ donation rate is increasing, with approximately 5,000 more recorded organ donations occurring in 2014 relative to 2006. To the right of the heatmap, next to each row, a bar displays the country's HDI ranking (a lower HDI ranking is better). Each country is colored based on which global region it belongs to: Europe (green), Eastern Mediterranean (purple), Western Pacific (yellow), America (orange), South East Asia (pink) and Africa (light green).

From Figure 2.5, we see that Spain is the clear leader in global organ donation, however there has been a rapid increase in donation rates in Croatia, which had one of the lower rates of organ donation in 2006 but has a rate equaling that of Spain in 2014. However, in contrast to the growth experienced by Croatia, the rate of organ donation appears to be slowing in several countries including as Germany, Slovakia and Cuba. For some unexplained reason, Iceland reported zero organ donations recorded from deceased donors in 2007.

The countries with the most organ donations are predominantly European and American. In addition, there appears to be a general correlation between organ donations and HDI ranking: countries with lower (better) HDI rankings tend to have higher organ donation rates. Subsequently, countries with higher (worse) HDI rankings tend to have lower organ donation rates, with the exception of a few Western Pacific countries such as Japan, Singapore and Korea, which have fairly good HDI rankings but relatively low organ donation rates.

In this case study, *superheat* allowed us to visualize multiple trends simultaneously without resorting to mass over-plotting. In particular, we were able to examine the organ donation over time and for each country and compare these trends to the country's HDI ranking while visually grouping countries from the same region together. No other 2-dimensional graph would be able to provide such an in-depth, yet uncluttered, summary of the trends contained in these data. In the next section, we will compare superheat to alternative graphs: parallel coordinates and scatterplot matrices.

The code used to produce Figure 2.5 is provided in the supplementary materials of [8].

A comparison with scatterplot matrices and parallel coordinates

The superheatmap in Figure 2.5 provides a clear, uncluttered view of this multivariate dataset (treating the primary variables in the data as the number of organ donations per year and the HDI ranking for each country). In this section, we will examine alternative views of the same data as presented by the two other popular multivariate plots: scatterplot matrices and parallel coordinate plots.

Figure 2.6 displays the organ donation data (originally presented in Figure 2.5) as a scatterplot matrix created using the *ggpairs* function from the *GGally* R package. Each row/column of the matrix corresponds to the number of organ donations for a given year and the final row/column corresponds to the HDI ranking. The scatterplot matrix presents each pair of variables as a scatterplot.

While the scatterplot matrix highlights the correlation between donation counts from one year to the next, unfortunately since we cannot follow a single country through time using these pairwise plots, it does not allow us to explore a time trend in the same way that the heatmap does.

Another alternative presentation of this data is in the form of a parallel coordinates plot. Instead of showing all countries on a single parallel coordinates plot, we show a tiling of parallel coordinate plots wherein each panel highlights a single country in Figure 2.7. Each country is represented by a line and each vertical axis represents a variable (the first 9 variables are the donations per year from 2006 to 2014, and the final variable is the HDI ranking). Each variable is scaled so that the bottom of each vertical axis line represents the smallest observed value for that variable and the top corresponds to the largest observed value. The parallel coordinates plot allows us to follow each country's donations over time and provides quite an effective representation of the donation trends over time for each country, as well as the country's performance relative to the other countries.

In both Figure 2.6 and Figure 2.7, the HDI ranking is presented as the same type of variable as the donations per year. As a result, the comparison of each country's organ donation trends with HDI is much less obvious than in the superheatmap from Figure 2.5. Moreover, the overall trend over time (the line plot above the heatmap in Figure 2.5) is absent from the parallel coordinate and scatterplot matrix versions.

In higher dimensional datasets, such as in our third case study in Section 2.6, neither a scatterplot matrix nor a parallel coordinates plot are appropriate due to unavoidable mass over-plotting. Thus, while all three of superheatmaps, scatterplot matrices, and parallel coordinates can be effective visualizations for data that reaches up to at most 50 dimensions, only the heatmap is able to handle datasets with hundreds or even thousands of rows or columns.

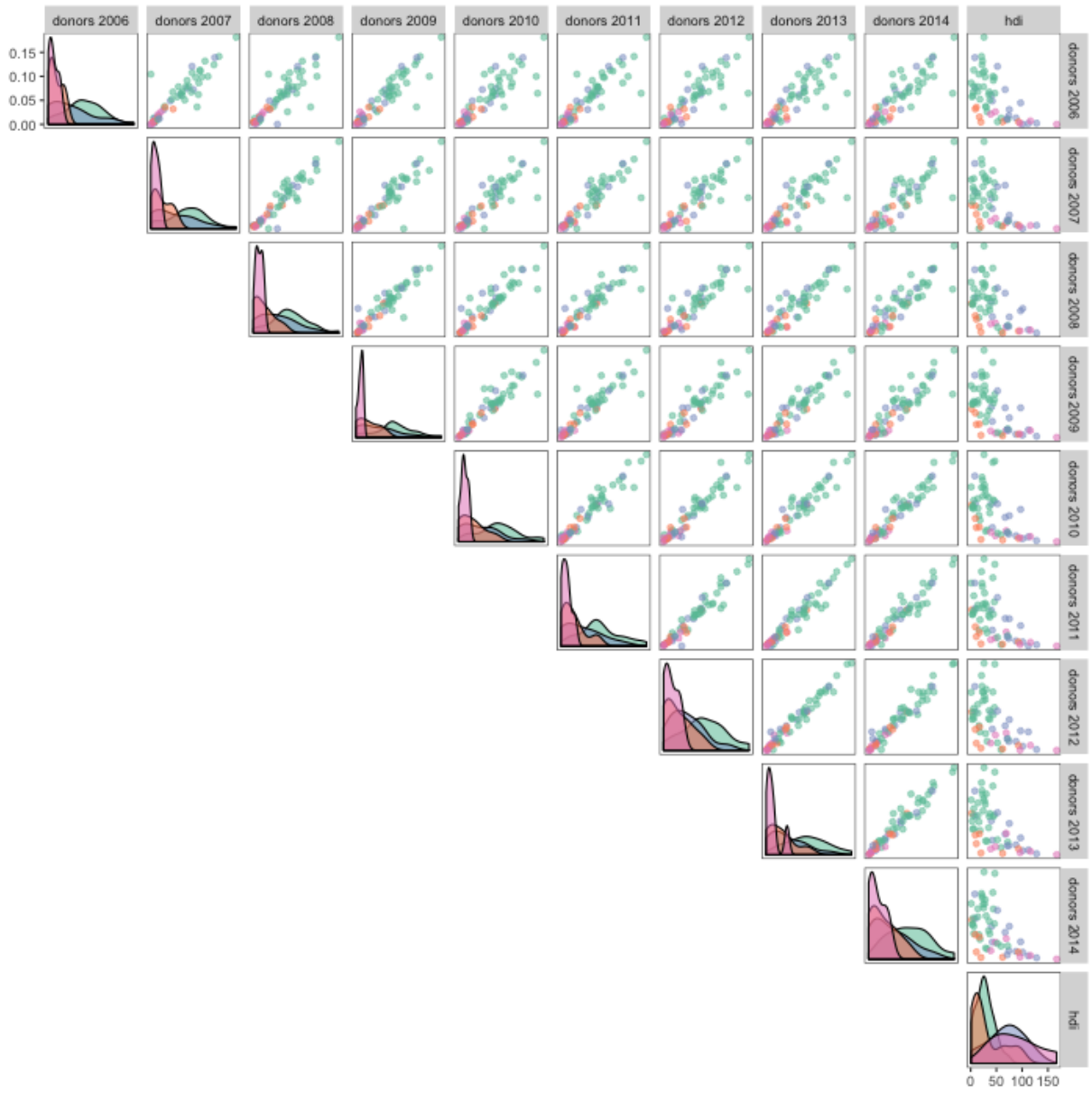

Figure 2.6: A scatterplot matrix of the organ donation data created using the *ggpairs* function from the *GGally* R package. The matrix contains of pairwise scatterplots for the following variables: the number of organ donations for each country each year from 2006 to 2014 and the country's HDI ranking. Each point is colored by region as in Figure 2.5.

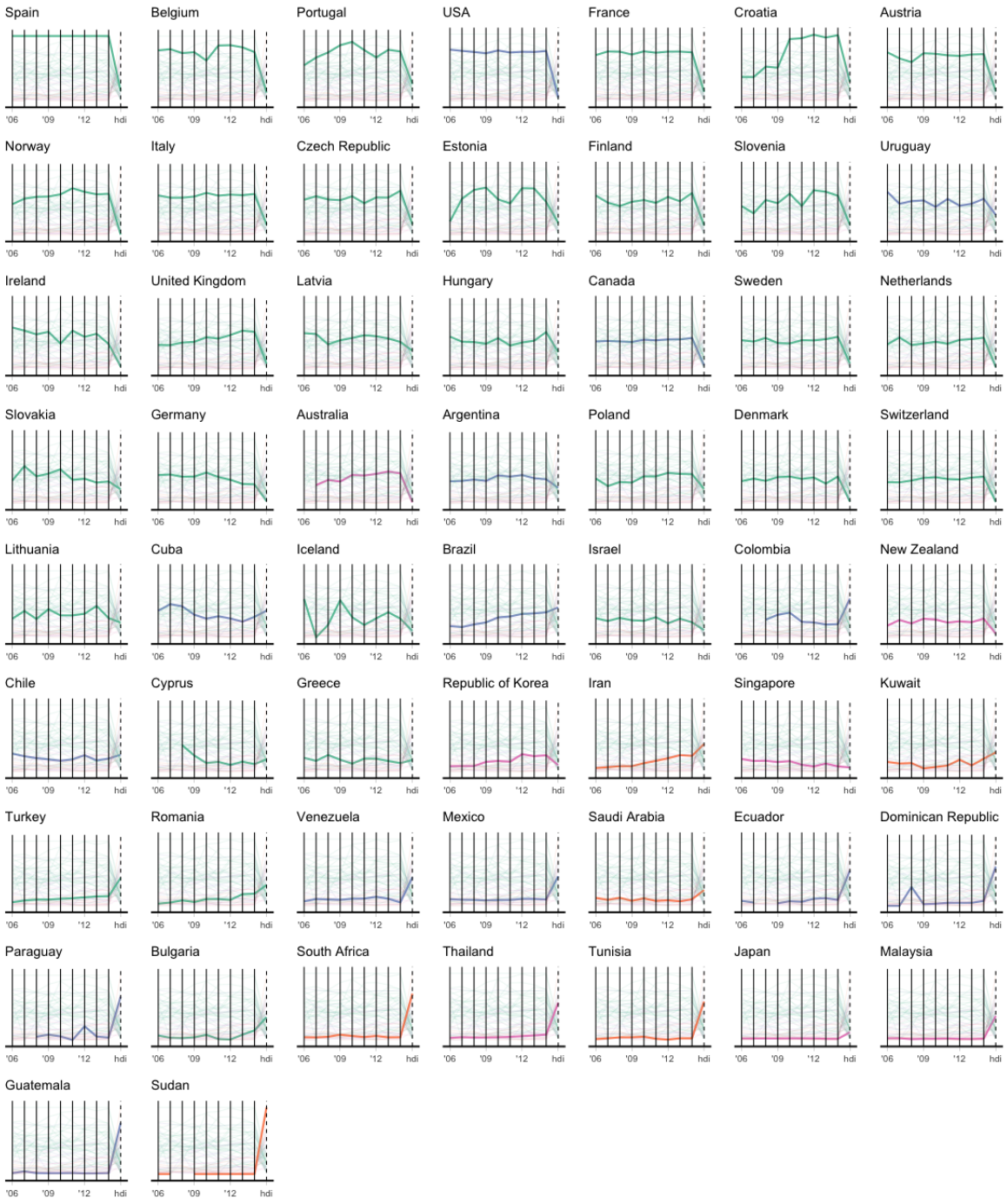

Figure 2.7: A series of parallel coordinates plots of the organ donation data built using the *ggplot2* R package. Each country corresponds to a line that traverses a path from one variable to another. Each variable has been scaled so that the bottom of the vertical line representing the variable corresponds to the smallest observed value and the top corresponds to the largest observed value. Each country is colored based on region as in Figure 2.5.

2.5 Case study II: uncovering clusters in language using Word2Vec

Word2Vec is an extremely popular group of algorithms for embedding words into high-dimensional spaces such that their relative distances to one another convey semantic meaning [69]. The canonical example highlighting the impressiveness of these word embeddings is

$$\overrightarrow{\text{man}} - \overrightarrow{\text{king}} + \overrightarrow{\text{woman}} = \overrightarrow{\text{queen}}.$$

That is, that if you take the word vector for "man", subtract the word vector for "king" and add the word vector for "woman", you approximately arrive at the word vector for "queen". These algorithms are quite remarkable and represent an exciting step towards teaching machines to understand language.

In 2013, Google published pre-trained vectors trained on part of the Google News corpus, which consists of around 100 billion words. Their algorithm produced 300-dimensional vectors for 3 million words and phrases [34].

The majority of existing visualization methods for word vectors focus on projecting the 300-dimensional space to a low-dimensional representation using methods such as t-distributed stochastic neighbor embedding (t-SNE) [63].

Visualizing cosine similarity

In this *superheat* case study we present an alternative approach to visualizing word vectors, which highlights contextual similarity. Figure 2.8 presents the cosine similarity matrix for the GoogleNews word vectors of the 35 most common words from the NY Times headlines dataset (from the *RTextTools* package). The rows and columns are ordered based on a hierarchical clustering and are accompanied by dendrograms describing this hierarchical cluster structure. From this superheatmap we observe that words appearing in global conflict contexts such as "terror" and "war" have high cosine similarity (implying that these words appear in similar contexts). Words that are used in legal contexts such as "court" and "case" as well as words with political context such as "Democrats" and "GOP" also have high pairwise cosine similarity. The code used to prepare Figure 2.8 is provided in the supplementary materials of [8].

Although the example presented in Figure 2.8 displays relatively few words (we are presenting only the 35 most frequent words) and we have reached our capacity to be able to visualize each word individually on a single page, it is possible to use *superheat* to represent hundreds or thousands of words simultaneously by aggregating over word clusters.

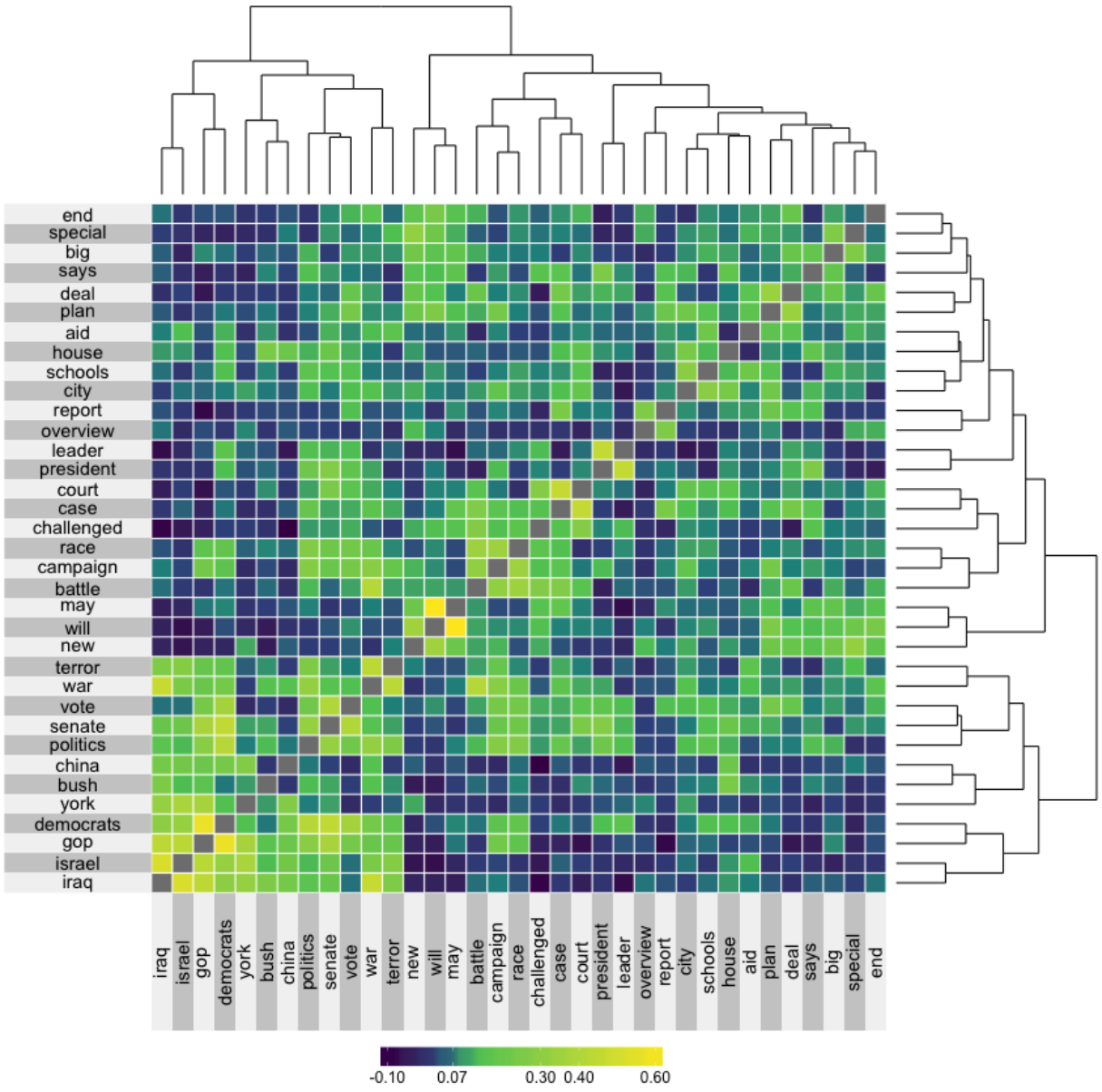

Figure 2.8: The cosine similarity matrix for the 35 most common words from the NY Times headlines that also appear in the Google News corpus. The rows and columns are ordered based on hierarchical clustering. This hierarchical clustering is displayed via dendrograms.

Visualizing word clusters

Figure 2.9(a) displays the cosine similarity matrix for the Google News word vectors of the 855 most common words from the NY Times headlines dataset where the words are grouped into 11 clusters generated using the Partitioning Around Medoids (PAM) algorithm [47, 82] applied to the rows/columns of the cosine similarity matrix. As PAM forces the cluster centroids to be data points, we represent each cluster by the word that corresponds to its center (these are the row and column labels that appear in Figure 2.9(a)). A silhouette plot is placed above the columns of the superheatmap in Figure 2.9(a), and the clusters are ordered in increasing average silhouette width.

The silhouette width is a traditional measure of cluster quality based on how well each object lies within its cluster, however we adapted its definition to suit cosine-based distance so that the cosine-silhouette width for data point i is defined to be:

$$sil_{\text{cosine}}(i) = b(i) - a(i)$$

where $a(i) = \frac{1}{\|C_i\|} \sum_{j \in C_i} d_{\text{cosine}}(x_i, x_j)$ is the average cosine-dissimilarity of i with all other data within the same cluster (C_i is the index set of the cluster to which i belongs), and $b(i) = \min_{C \neq C_i} d_{\text{cosine}}(x_i, C)$ is the lowest average dissimilarity of i to any other cluster of which i is not a member. $d_{\text{cosine}}(x, y)$ is a measure of cosine "distance", which is equal to $d_{\text{cosine}} = \frac{cos^{-1}(s_{\text{cosine}})}{\pi}$ (where s_{cosine} is standard cosine similarity).

The number of clusters ($k = 11$) was chosen based on the value of k that was optimal under two criteria: (1) performance-based [84]: the maximal average cosine-silhouette width, and (2) stability-based [114]: the average pairwise Jaccard similarity based on 100 membership vectors each generated by a 90% subsample of the data. Plots of k versus average silhouette width and average Jaccard similarity are presented in Appendix Figure A.1.

Word clouds displaying the words that are members of each of the 11 word clusters are presented in Appendix Figure A.2. For example, the "government" cluster contains words that typically appear in political contexts such as "president", "leader", and "senate", whereas the "murder" cluster contains words such as "case", "drugs", and "crime".

Figure 2.9(b) presents a "smoothed" version of the cosine similarity matrix in panel (a), where the smoothed cluster-aggregated value corresponds to the median of the original values in the "un-smoothed" matrix. The smoothing provides an aggregated representation of Figure 2.9(a) allowing the viewer to focus on the overall differences between the clusters. Note that the color range is slightly different between panels (a) and (b) due to the extreme values present in panel (a) being removed when we take the median in panel (b).

What we find is that the words in the "American" cluster have high silhouette widths, and thus is a "tight" cluster. This is reflected in the high cosine similarity within the cluster and low similarity between the words in the "American" cluster and words from other clusters. However, the words in the "murder" cluster have relatively high cosine similarity with words in the "government", "struggle", and "bombing" clusters.

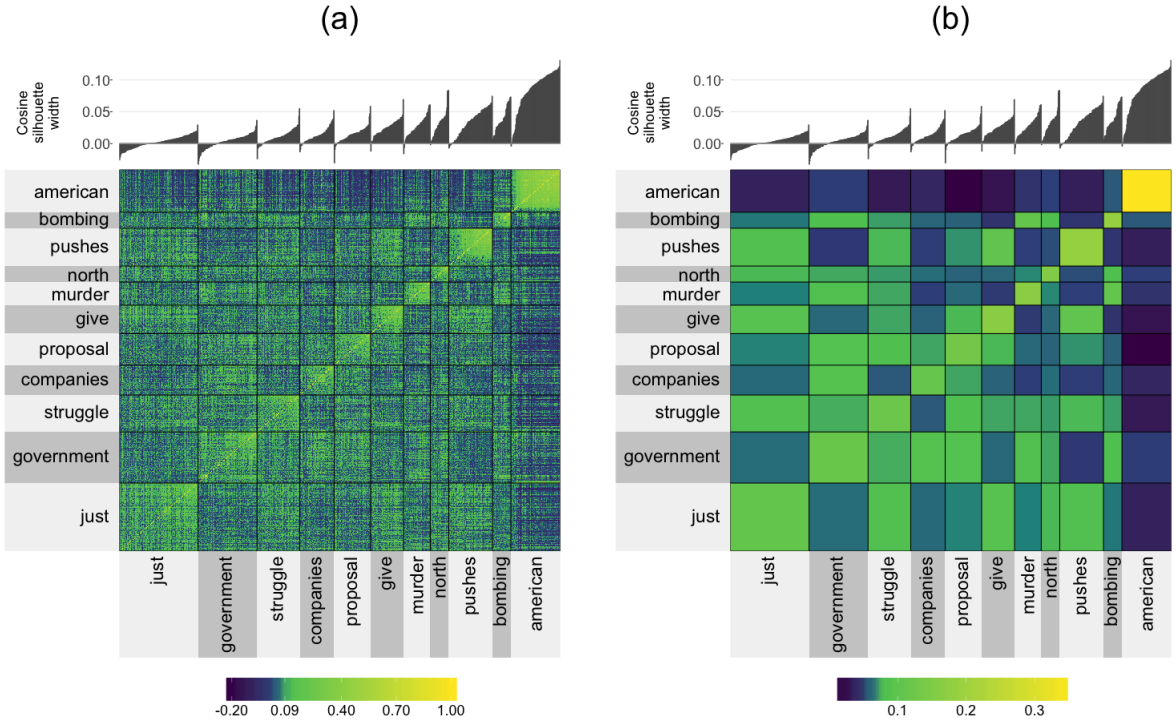

Figure 2.9: A clustered cosine similarity matrix for the 855 most common words from the NY Times headlines that also appear in the Google News corpus. The clusters were generated using PAM and the cluster label is given by the medoid word of the cluster. Panel (a) displays the raw clustered 855×855 cosine similarity matrix, while panel (b) displays a "smoothed" version where the cells in the cluster are aggregated by taking the median of the values within the cluster.

The clusters whose centers are not topic-specific such as "just" and "pushes" tend to consist of common words that are context agnostic (see their word clouds in Appendix A.2), and these clusters have fairly high average similarity with one another.

The information presented by Figure 2.9 far surpasses that of a standard silhouette plot: it allows the quality of the clusters to be evaluated relative to one another. For example, when a cluster exhibits low between-cluster separability, we can clearly see *which* clusters it is close to.

The code used to produce Figure 2.9 is provided in the supplementary materials of [8].

2.6 Case study III: evaluation of heterogeneity in the performance of predictive models for fMRI brain signals from image inputs

Our final case study evaluates the performance of a number of models of the brain's response to visual stimuli. This study is based on data collected from a functional Magnetic Resonance Imaging (fMRI) experiment performed on a single individual by the Gallant neuroscience lab at UC Berkeley [103, 102].

fMRI measures oxygenated blood flow in the brain, which can be considered as an indirect measure of neural activity (the two processes are highly correlated). The measurements obtained from an fMRI experiment correspond to the aggregated response of hundreds of thousands of neurons within cube-like voxels of the brain, where the segmentation of the brain into 3D voxels is analogous to the segmentation of an image into 2D pixels.

The data contains the fMRI measurements (averaged over 10 runs of the experiment) for each of 1,294 voxels located in the V1 region of the visual cortex of a single individual in response to viewings of 1,750 different images (such as a picture of a baby, a house or a horse). Each image is a 128×128 pixel grayscale image, which is represented as a vector of length $10,921$ through a Gabor wavelet transformation [58]. Figure 2.10 displays a graphical representation of the data structure.

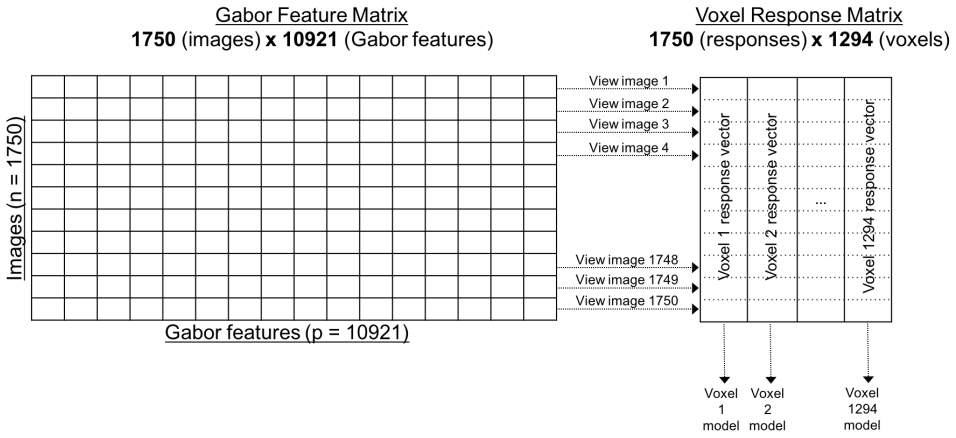

Figure 2.10: A diagram describing the fMRI data: a design matrix with 1,750 observations (images) and 10,921 features (Gabor wavelets) for each image, and a voxel response matrix consisting of 1,294 distinct voxel response vectors, where, for each voxel, the responses to each of the 1,750 images were collected. We fit a predictive model for each voxel using the Gabor feature matrix (1,294 models). The heatmap in Figure 2.11 corresponds to the voxel response matrix.

Data access can be located in [48]. However, unfortunately, only the voxel responses and raw images are available. The Gabor wavelet features are not provided.

Modeling brain activity

We developed a model for each voxel that predicts its response to visual stimuli in the form of greyscale images. Since each voxel responds quite differently to the image stimuli, instead of fitting a single multi-response model, we fit 1,294 independent Lasso models as in [102].

The models are then evaluated based on how well they predict the voxel responses to a set of 120 withheld validation images.

Simultaneous performance evaluation of all 1,294 voxel-models

The voxel response matrix is displayed in Figure 2.11(a). The rows of the heatmap correspond to the 120 images from the validation set, while the columns correspond to the 1,294 voxels. Each cell displays the voxel's response to the image. The rows and columns are clustered into two groups using K-means. As in Figure 2.9(a), the heatmap is extremely grainy. Figure 2.11(b) displays the same heatmap with the cell values smoothed within each cluster (by taking the median value).

Figures A.4 and A.3 in the Appendix display four randomly selected images from each of the two image clusters. We find that the bottom image cluster consists of images for which the subject is easily identifiable (e.g. Princess Diana and Prince Charles riding in a carriage, a bird, or an insect), whereas the contents of images from the top cluster of images are less easy to identify (e.g. rocks, a bunch of apples, or an abstract painting).

From Figure 2.11, it is clear that the brain is much more active in response to the images from the top cluster (whose contents were less easily identifiable) than to images from the bottom cluster.

Furthermore, there are two distinct groups of voxels:

1. **Sensitive voxels** that respond very differently to the two groups of images (for the top image cluster, their response is significantly lower than the average response, while for the bottom image cluster, their response is significantly higher than the average response).

2. **Neutral voxels** that respond similarly to both clusters of images.

In addition, above each voxel (column) in the heatmap, the correlation of that voxel-model's predicted responses with the voxel's true response is presented (as a scatterplot in Panel (a) and as aggregate boxplots in Panel (b)).

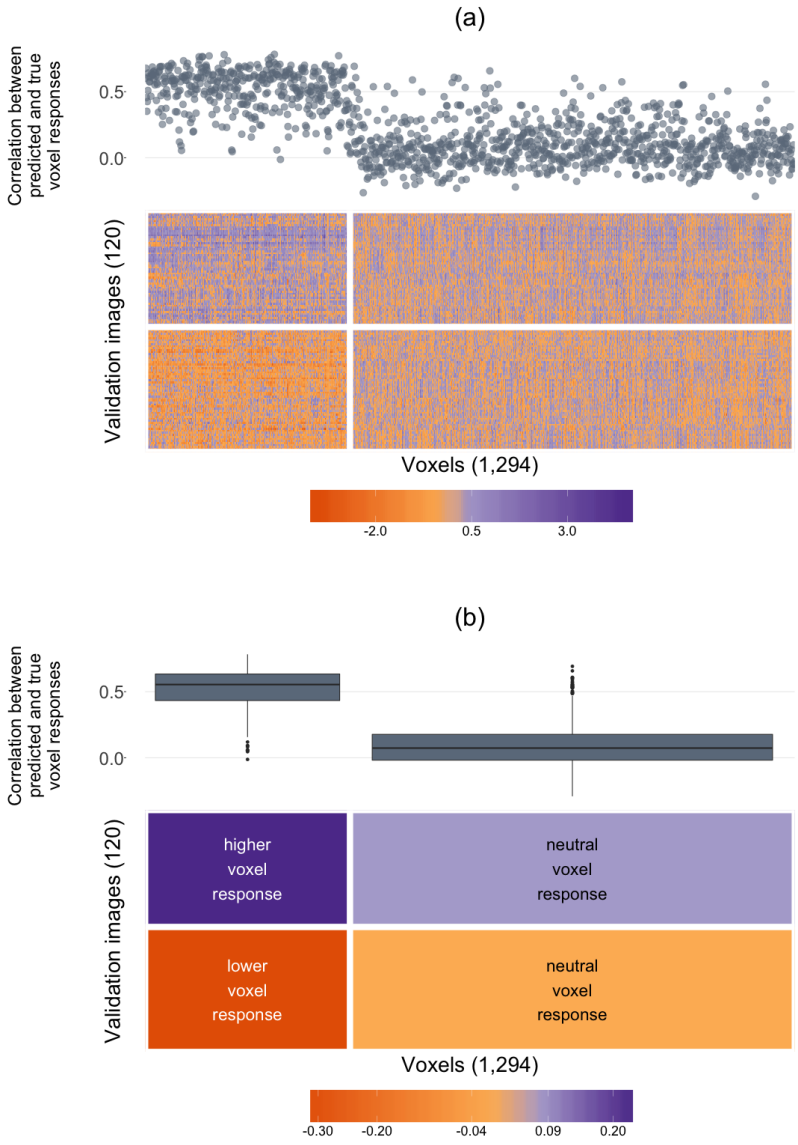

Figure 2.11: A superheatmap displaying the validation set voxel response matrix (Panel (a) displays the raw matrix, while Panel (b) displays a smoothed version). The images (rows) and voxels (columns) are each clustered into two groups (using K-means). The left cluster of voxels are more "sensitive" wherein their response is different for each group of images (higher than the average response for top cluster images, and lower than the average response for bottom cluster images), while the right cluster of voxels are more "neutral" wherein their response is similar for both image clusters. Voxel-specific Lasso model performance is plotted as correlations above the columns of the heatmap (as a scatterplot in (a) and cluster-aggregated boxplots in (b)).

It is clear that the models for the voxels in the first (sensitive) cluster perform significantly better than the models for the voxels in the second (neutral) cluster. That is, the responses of the voxels that are sensitive to the image stimuli are much easier to predict (the average correlation between the predicted and true responses was greater than 0.5) than the responses of the voxels whose responses are neutral (the average correlation between the predicted and true responses was close to zero).

Further examination revealed that the neutral voxels were primarily located on the periphery of the V1 region of the visual cortex, whereas the sensitive voxels tended to be more centrally located.

Although a standard histogram of the predicted and observed response correlations would have revealed that there were two groups of voxels (those whose responses we can predict well, and those whose responses we cannot), *superheat* allowed us to examine this finding in context. In particular, it allowed us to take advantage of the heterogeneity present in the data: we were able to identify that the voxels whose response we were able to predict well were exactly the voxels whose response was sensitive to the two clusters of images.

Note that we also ran Random Forest models for predicting the voxel responses and found the same results, however, the overall correlation was approximately 0.05 higher on average.

The code used to produce Figure 2.11 is provided in the supplementary materials of [8].

2.7 Conclusion

In this Chapter, we have proposed the superheatmap that augments traditional heatmaps via the inclusion of extra information such as a response variable as a scatterplot, model results as boxplots, correlation information as barplots, text information, and more. These augmentations provide the user with an additional avenue for information extraction, and allow for exploration of heterogeneity within the data. The superheatmap, as implemented by the *superheat* package written by the authors, is highly customizable and can be used effectively in a wide range of situations in exploratory data analysis and model assessment. The usefulness of the superheatmap was highlighted in three case studies. The first combined multiple sources of data to assess the relationship between organ donation and country development worldwide. The second explored the structure of the English language by visualizing word clusters from Word2Vec data, while highlighting the hierarchical nature of these word groupings. Finally, the third case study evaluated heterogeneity in the performance of Lasso models designed to predict fMRI brain signals in response to visual stimuli in the form of image viewings. We hope that we have demonstrated clearly that the heatmap is an extremely useful data visualization tool, particularly for high-dimensional datasets.

Part II

Prediction: Predicting Surgical Site Infections Using Electronic Medical Records

Chapter 3

A History of Predicting Surgical Site Infections

3.1 Introduction

While all surgeries have inherent risks, one of the most serious risks is Surgical Site Infection (SSI), defined as a post-operative infection that forms at the site of the surgery within 30 days of the procedure. While SSI only occurs in approximately 2-5% of all surgeries (depending on the procedure being undertaken and the hospital at which it is undertaken), it is responsible for up to 30% of all Hospital Acquired Infections (HAI) [64, 2]. SSI contributes to increased morbidity, mortality, poor quality of life, prolonged hospital stay, and increased readmission rate and healthcare expenditure [24, 57]. Identifying patients who are at risk for SSI and implementing preventative measures has the potential to drastically improve surgical outcomes in hospitals worldwide. However, even though understanding of the features associated with SSI has progressed over the past decade, a simple, accurate, and openly available hospital-agnostic model that can be used to predict SSI does not yet exist [24]. The predictive models that are used were built in the 1990s and early 2000s and do not represent the capabilities of today's technological era.

With the current era's vast amounts of accessible healthcare data, together with recent advances in machine learning methods and computational power, we are finally in a position to be able to develop models for accurately predicting patients who are at increased risk for SSI by combining mandatory SSI surveillance data with internal patient data routinely collected by almost every hospital. While there do exist a few studies aiming to develop such models, the existing approaches are hampered by small sample sizes and unrealistic cohorts that do not reflect reality. Instead, the majority of work in this arena is focused on identifying risk factors for SSI, rather than directly predicting SSI.

In Part II of this thesis, we formulate a solution to the problem of identifying patients at risk of SSI. If it is possible to identify patients at risk of SSI, then the task of reducing SSI in hospitals is substantially easier since it allows for closer monitoring and subsequently

early intervention.

To achieve this goal, we first collect, explore, and clean a relevant dataset consisting of SSI surveillance data and electronic medical records from UC Davis. Using this data, we develop a simple, intuitive model for predicting a patient's risk of developing SSI. Using such a model, clinicians can be alerted to the risk and use this information to make informed decisions about SSI prevention. All of the input features for our modeling approach are readily available from routinely collected Electronic Health Record (EHR) data, providing an opportunity to automate the process of prediction and providing decision support at the point-of-care. Such an automated prediction pipeline has the potential to have a broad impact across the continuum of pre-operative surgical planning phase into the post-operative recovery phase.

The remainder of this chapter is structured as follows. Section 3.2 describes the risk associated with SSI and the surveillance methods and data collection protocols currently in place. Section 3.3 formulates the problem we are trying to solve. Section 3.4 introduces and explores the data from UC Davis that we will use in the next chapter to solve the problem. Finally, Section 3.5 describes the preprocessing steps we undertook with the data to prepare it for modelling.

3.2 Surgical site infections surveillance initiatives

The impact of SSI on patients, hospitals, and public health in general, are enormous. Along with pain, discomfort, and the need for additional interventions, SSIs are estimated to add an additional 7 to 11 days to patients' length of hospital stay; a 2- to 11-fold increase in risk of death compared to non-SSI patients; and 77% of deaths among patients with SSI are directly linked to the SSI [64, 2].

Since SSI has proven to be a critical patient outcome, it is among one of the key quality measures that are used to compare hospitals across the USA, and subsequently, hospital insurance reimbursement payments are dependent on the hospital's SSI rates. Further, many of the costs associated with managing SSI are non-reimbursable. According to one cost estimate, SSIs add about $3.5 billion to $10 billion annually to the healthcare expenditures [3]. SSI are a national healthcare priority and several initiatives for SSI surveillance within hospitals have arisen over the past few decades, with the goal of improving early detection of SSI [79].

Early SSI surveillance efforts by the Centers for Disease Control and Prevention (CDC) introduced programs for tracking SSI rates and other surgical outcomes over time and across institutions. These programs mandated (with a pay-for-performance incentive) that hospitals across the country enter their SSI-related data into the National Healthcare Safety Network (NHSN) database. In addition to the NHSN database, The American College of Surgeons later implemented another surveillance database called the National Surgical Quality Improvement Program (NSQIP), participation in which is voluntary.

The data collection methods at point-of-care for both NHSN and NSQIP are active surveillance and retrospective review of clinical documentation, though the NSQIP surveillance methods differ slightly from those of the NHSN. Both NHSN and NSQIP standards, definitions, and surveillance data are broadly accepted benchmarks that are implemented across healthcare facilities as part of quality improvement initiatives and surgical outcomes research. Both data repositories provide manually curated and validated data from both SSI and non-SSI control populations from healthcare facilities all across US [11]. However the data collected by NSQIP represents a random sample of surgeries, whereas the data collected by NHSN represent a more complete view of the surgeries performed at any single hospital.

In this thesis we will use the EHR records and NHSN database from the UC Davis School of Medicine from 2014 to 2017 to develop a predictive model for SSI that can be used both before and after surgery.

3.3 Formulating the SSI prediction problem

While our overall goal is to reduce the rate of SSI in hospitals, in this thesis, we formulate a specific prediction-based problem that will allow for concrete steps to be taken towards achieving this goal. Specifically, we will develop an approach for predicting if a patient is at risk of SSI, allowing clinicians to implement timely interventions, and increased monitoring to both decrease the impact of SSIs on patients, as well as reduce the overall rate of SSI. However, before we are ready to apply algorithms to our data, we need to properly formulate our problem based on the data available.

Since we will be using the UC Davis NHSN SSI surveillance database, our problem specification will need to be based on the NHSN definition of an SSI. This means that a patient is classified as having an SSI if they have an infection at the site of surgery within the 30 days following the surgical procedure. While NHSN technically has three separate SSI classifications: (1) superficial incisional SSI (skin and subcutaneous tissue-level), (2) deep incisional SSI (deep soft tissue-level), (3) and organ/space SSI (any part of the body that is deeper than the fascial/muscle layers involved in the operative procedure), to keep our problem manageable, we make no such distinction between these classifications. For our problem, a patient either has SSI (which could be any of the three types), or they do not.

Moreover, since around 30% of patients have multiple procedures recorded in the database, we need to decide whether to define an observational unit as a single procedure, or as a single patient. Since a patient might have an infection for some procedures, but not others, we decided to treat each observational unit as a single procedure, which means that our final covariate matrix (which has one row per observational unit) has multiple rows for some patients corresponding to their multiple procedures.

As our dataset involves combining the NHSN SSI surveillance data with EHR data, we also needed to decide whether to join by procedure ID or by patient ID. Since we found that many EHR lab, vitals, and medication data did not have matching procedure IDs with the

NHSN data, but did have matching patient IDs, we decided to join the EHR and NHSN data using patient ID and recorded date, rather than procedure ID.

To further refine our problem statement to "predict if a patient is at risk of SSI", we need to specify what we mean by "at risk". In the end, our analytic process led us to an integer risk scoring system. Initially, our algorithm predicts a value in between 0 and 1 corresponding to the average predicted probability of SSI across many models. However, since this value itself should not be interpreted literally as a probability, we converted this average predicted value to an integer score between 0 and 10, which corresponds to the average predicted probability value multiplied by 10, and rounded to the nearest integer. This integer score is what we call the "SSI score". The higher the SSI score, the higher the estimated risk of SSI.

Next, we had to decide whether to develop a single predictive model, or separate models for each surgical procedure. However, since the number of SSI cases in our data for most of the individual surgical procedures was fewer than 10, we determined that we did not have enough data to consider each procedure separately. Based on this idea, we also considered grouping procedures together where the groups were based both on how similar procedures look to be in the data, as well as based on the expert opinions of our collaborators. However, again we found that for many of the procedure groups, the number of SSI cases was too small to draw any substantial conclusions about predictive accuracy. In the end, we settled on a universal model for all procedures, and we will show that this model is more accurate than the separate grouped procedure models. Thus our final problem involves generating a procedure-agnostic SSI integer score between 0 and 10 corresponding to SSI risk for each new patient.

These analytic decisions were each made to ensure that we were capturing the most complete and relevant subset of the data for addressing our prediction problem, as well as based on discussion with our medical domain collaborators at UC Davis.

The next section describes and explores the UC Davis NHSN and the EHR datasets.

3.4 The UC Davis NHSN and EHR data

To develop an approach for predicting surgical site infections, we extracted all of the NHSN data and hospital EPIC Electronic Health Record (EHR) data from all surgeries undertaken between 2014 and 2017 at the University of California Davis School of Medicine. Together, the UC Davis NHSN and EHR databases consisted of 6 different sources of data spread across 24 separate files for 2014 through to 2017. The 6 sources of data corresponded to

- **NHSN "numerator" SSI data** with 936 rows and 139 variables. This dataset contains information on the patients who were diagnosed with SSI (the "numerators"). The variables include information on the procedure such as estimated blood loss, the pathogen or organism identified in the infection, whether the patient died, etc.

- **NHSN "denominator" surgery data** with 39,174 rows and 44 variables for all patients who underwent a surgery, whether or not they got an SSI (the "denominators"). The variables include the time of the surgery, the age of the surgeon, whether the surgery was laparascopic, whether the patient was given anesthesia, their ASA health status prior to the surgery, whether the surgery was inpatient or outpatient, etc. This dataset also contains information about the patient including the patient's age, gender, BMI, etc.

- **Lab EHR data** with 12,927,273 rows and 30 lab variables. These lab variables include alanine transferase (ALT), albumin, alkaline phosphatase (ALP), aspartate transaminase (AST), basophils ABS, bilirubin, C-Reactive Protein (CRP), calcium, carbon dioxide, chloride, creatinine serum, E-GFR, eosinophil count, glucose, hematocrit, hemoglobin, lymphocytes, monocytes ABS, neutrophil ABS, platelet count, potassium, protein, red cell count, sodium, urea nitrogen, and white blood cell count.

- **Vitals EHR data** with 8,666,375 rows on 9 variables involving recent and historical measurements on temperature, pulse, weight, height, and BMI.

- **Medication prescription EHR data** 7,637,621 rows and 50 lab therapeutic class category variables. This dataset contains all medication classes prescribed to each patient both historically and related to the surgery.

There are a total of 30,791 unique procedures performed on 27,326 patients represented in the data. 2.5% of these procedures had an associated SSI event. There are 38 types of procedures captured in the data, and the most common procedures were fracture (`FX`) with 3,300 cases, exploratory abdominal (`XLAP`) with 2,843 cases, and herniorrhaphy (`HER`) with 2,469 cases.

The remainder of this section summarizes and explores each of these datasets.

NHSN "numerator" data on SSI surgeries

Since the variables collected in this spreadsheet are only available for the SSI patients (but not for the non-SSI patients), this data was only used to identify which patients were diagnosed with SSI, and also to identify whether they had an infection already present at the time of surgery (in which case they were excluded from our cohort).

NHSN "denominator" data on all surgeries

The "denominator" data collected by the NHSN is the main dataset containing information on each patient and their surgery, which in our case includes all surgeries that took place at UC Davis between 2013-12-20 to 2017-10-13.

We will examine the data by considering a single randomly selected patient whose patient ID (`PATNUM`) is 33086929. This patient is a 64-year-old white male who underwent prostate

surgery on March 20 2015. In this data file, there is a single row for each surgery, so patients who had multiple surgeries have multiple rows in the data. 7,680 patients out of a total of 27,326 patients had more than one surgery (and thus appear in more than one row in the data) with different procedure IDs (PROCIDs). Table 3.1 provides a list of all variables in this denominator dataset, a description of each variable, as well as the value reported for patient 33086929.

Since there are many different procedures represented in the data, Table 3.2 displays the procedures in the data arranged in decreasing order of prevalence. This table also contains information provided by our surgeon collaborators on the risk of each procedure, the proportion of the procedures performed on women, the proportion of patients under anesthesia during the procedure, the proportion of procedures that were outpatient procedures, and the average length of the surgery.

Variable	Description	Value for PATNUM 33086929
PATNUM	De-identified Patient ID	33086929
ADMISSION_ENCNUM	De-identified surgical encounter identifier	915571140
GENDER	Patient's gender	M
PAT_PROC_AGE	Age at time of surgery	64
PROCID	Unique record/row ID	19349983
PROCDATE_SET	Date of surgery	20-3-15
PROCCODE	NHSN Surgical procedure	PRST
ANESTHESIA	Was anesthesia administered?	Y
ASA	ASA physical condition status	3
CLOSURE	Incisional would closure type	PRIMARY
EMERGENCY	Procedure an emergency or urgent procedure?	N
OUTPATIENT	Patient discharged on day of admission?	N
RISK	Risk level of the procedure	1
SCOPE	Was the procedure laparoscopic?	Y
SWCLASS	Surgical wound class	CLEAN
TRAUMA	Was there a blunt or penetrating injury?	N
CASE_ID	The case identifier for the surgery	28309
IN_OP_ROOM_SET	The date/time patient entered the OR	20-3-15 16:43:00
OUT_OF_ROOM_SET	The date/time patient departed the OR	20-3-15 20:13:00
SURGEON_CODE	Code of the surgeon	70
SURGICAL_SERVICE	The surgical service that performed the surgery	Urology
PRIMARY_DX	The Principal ICD10 diagnosis code	185
PRIMARY_DX_DESC	The Principal ICD10 diagnosis name	Malignant neoplasm of prostate
PATIENT_RACE	Patient-declared race	White
SMOKING_ASSESSMENT_DATE	Date of most recent smoking status	17-3-15
SMOKING_STATUS	Patient-declared smoking status	FORMER SMOKER
RBC_TRANSFUSED	The no. of red blood cell units transfused	NA
PLATELETS_TRANSFUSED	The no. of platelet units transfused	NA
FFP_TRANSFUSED	The no. of fresh frozen plasma units transfused	NA
CRYO_TRANSFUSED	The no. of cryoprecipitate units transfused	NA
INCARCERATED	Whether the surgical patient was incarcerated	NA

Table 3.1: The list of variables in the denominator data, with an example value from patient 33086929.

PROCCODE	Description	SSI Risk	N	Female (%)	Anesthesia (%)	Outpatient (%)	Average surgery length (hours)
FX	Fracture	Low	3300	42	97	14	3.9
XLAP	Exploratory abdominal	High	2843	55	98	6	3.9
HER	Herniorrhaphy	High	2469	22	99	69	2.5
LAM	Laminectomy	Low	2097	46	99	11	5.0
CSEC	Cesarean section	Med	2007	100	9	0	1.9
FUSN	Spinal fusion	Low	1957	49	99	0	6.2
CHOL	Gallbladder	High	1656	68	100	40	2.8
BRST	Breast	Low	1620	97	95	73	2.5
CRAN	Craniotomy	Low	1514	44	94	0	4.8
NEPH	Kidney surgery	Med	1308	44	99	4	4.1
HPRO	Hip prosthesis	Low	1306	56	70	0	3.4
OVRY	Ovarian	Med	1263	100	93	33	3.4
THOR	Thoracic	Med	1175	44	98	4	4.1
SB	Small bowel	High	1149	44	99	1	4.7
KTP	Kidney transplant	High	1142	41	99	0	5.1
KPRO	Knee prosthesis	Low	1092	61	51	0	3.2
APPY	Appendix	High	1072	46	100	6	2.1
CARD	Cardiac	Low	1005	35	99	0	6.8
COLO	Colon	High	994	45	100	1	4.8
GAST	Gastric	Med	987	62	99	2	4.0
HYST	Abdominal hyst.	Med	978	100	98	19	4.5
BILI	Bile duct liver pancr.	High	918	63	99	3	4.0
AMP	Limb amputation	Med	711	28	79	5	2.4
THYR	Thyroid	Low	707	72	100	22	4.0
VSHN	Ventricular shunt	Low	688	45	99	2	2.9
AVSD	AV shunt dialysis	Low	493	48	39	77	2.3
NECK	Neck	Med	416	32	95	12	7.1
PRST	Prostate	Med	413	0	99	4	4.1
CBGB	Coronary bypass donor	Low	359	18	99	0	8.2
PACE	Pacemaker	Med	320	40	53	15	4.3
REC	Rectal	High	261	48	95	8	5.7
SPLE	Spleen	Med	215	46	100	0	4.0
VHYS	Vaginal hysterectomy	Med	213	100	100	44	3.8
PVBY	Peripheral vasc. bypass	Low	199	30	96	1	8.1
CEA	Carotid endarterectomy	Low	132	34	99	0	4.5
RFUSN	Refusion spine	Low	100	59	99	0	8.0
CBGC	Coronary bypass graft	Low	51	18	100	0	7.9
AAA	Aortic aneurysm	Med	44	41	100	0	8.1

Table 3.2: A summary of the 38 procedures.

The boxplots in Figure 3.1 show that SSI patients have slightly longer surgeries in general than the non-SSI patients, but there is very little difference in BMI and age between SSI and non-SSI patients (if anything, the SSI patients are slightly older).

The dot plot in Figure 3.2 show the difference in the proportion of patients with the positive class of each categorical variable for the SSI and non-SSI classes. Substantially more SSI patients have RBCs transfused than non-SSI patients, and more non-SSI patients have a clean surgical wound classification than SSI patients.

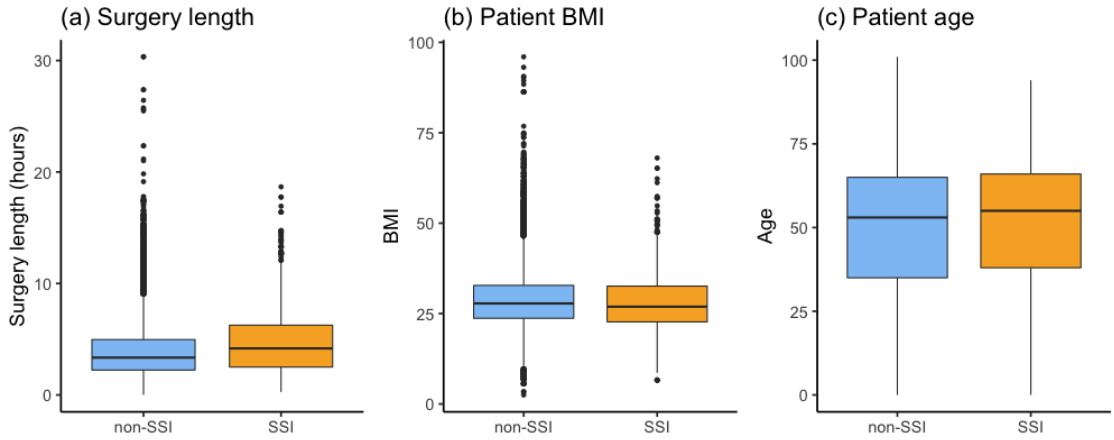

Figure 3.1: Boxplots displaying the distribution of (a) surgery length, (b) BMI, and (c) age for the non-SSI and SSI patients.

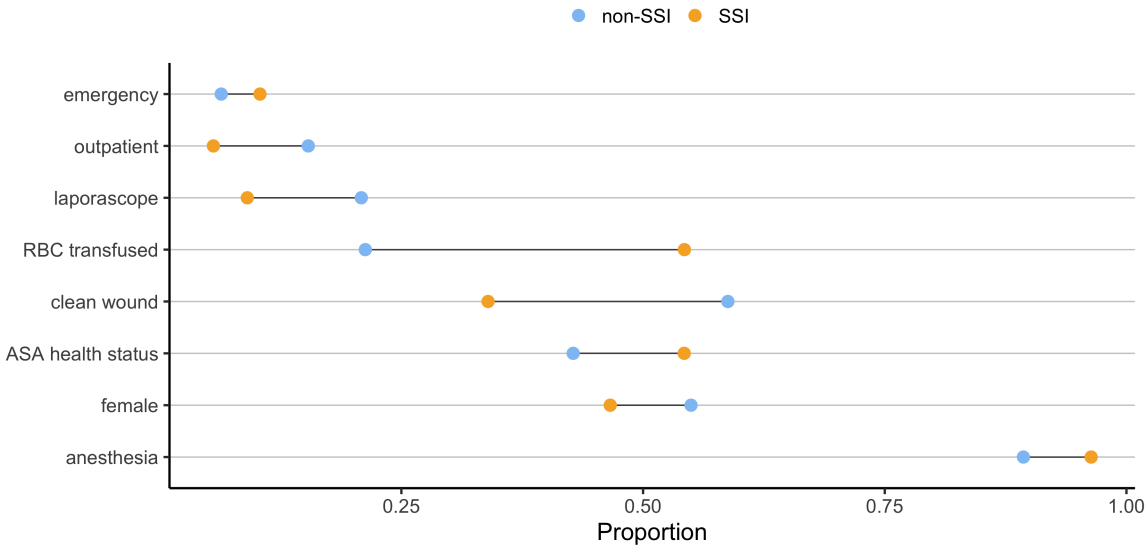

Figure 3.2: Dot plots displaying the proportion of patients with the positive class for each binary variable separated by SSI category.

Lab EHR data

The lab data was provided in a long-format where every individual measurement taken had it's own row as in 3.3. A description of each lab measurement is provided in Table A.2.

PATNUM	ENCNUM	DATE	NAME	NUM_VALUE	ORD_VALUE
33086929	915571140	20-3-15 21:13:00	E-GFR, AA	9999999.00	> 60
33086929	915571140	20-3-15 21:13:00	GLUCOSE	151.00	151
33086929	915571140	20-3-15 21:13:00	PLATELET COUNT	222.00	222
33086929	915571140	20-3-15 21:13:00	E-GFR, NON-AA	9999999.00	> 60
33086929	915571140	20-3-15 21:13:00	UREA NITROGEN	12.00	12
33086929	915571140	20-3-15 21:13:00	HEMOGLOBIN	13.10	13.1
33086929	915571140	20-3-15 21:13:00	RED CELL COUNT	4.14	4.14
33086929	915571140	20-3-15 21:13:00	POTASSIUM	4.40	4.4
33086929	915571140	20-3-15 21:13:00	SODIUM	136.00	136
33086929	915571140	20-3-15 21:13:00	CALCIUM	8.50	8.5
33086929	915571140	20-3-15 21:13:00	HEMATOCRIT	38.90	38.9
33086929	915571140	20-3-15 21:13:00	CARBON DIOXIDE TOTAL	24.00	24
33086929	915571140	20-3-15 21:13:00	CHLORIDE	103.00	103
33086929	915571140	20-3-15 21:13:00	MONOCYTES ABS AUTO	0.70	0.7
33086929	915571140	20-3-15 21:13:00	WHITE BLOOD CELL COUNT	14.10	14.1
33086929	915571140	20-3-15 21:13:00	NEUTROPHIL ABS AUTO	12.40	12.40
33086929	915571140	21-3-15 03:10:00	GLUCOSE	124.00	124
33086929	915571140	21-3-15 03:10:00	E-GFR, NON-AA	9999999.00	> 60
33086929	915571140	21-3-15 03:10:00	UREA NITROGEN	11.00	11
33086929	915571140	21-3-15 03:10:00	POTASSIUM	4.60	4.6
33086929	915571140	21-3-15 03:10:00	CALCIUM	8.60	8.6
33086929	915571140	21-3-15 03:10:00	HEMOGLOBIN	13.70	13.7
33086929	915571140	21-3-15 03:10:00	PLATELET COUNT	228.00	228
33086929	915571140	21-3-15 03:10:00	E-GFR, AA	9999999.00	> 60
33086929	915571140	21-3-15 03:10:00	CHLORIDE	100.00	100
33086929	915571140	21-3-15 03:10:00	CARBON DIOXIDE TOTAL	27.00	27
33086929	915571140	21-3-15 03:10:00	HEMATOCRIT	40.80	40.8
33086929	915571140	21-3-15 03:10:00	WHITE BLOOD CELL COUNT	8.70	8.7
33086929	915571140	21-3-15 03:10:00	SODIUM	135.00	135
33086929	915571140	21-3-15 03:10:00	RED CELL COUNT	4.36	4.36
33086929	933658868	14-3-15 12:02:00	RED CELL COUNT	4.46	4.46
33086929	933658868	14-3-15 12:02:00	CARBON DIOXIDE TOTAL	25.00	25
33086929	933658868	14-3-15 12:02:00	POTASSIUM	4.30	4.3
33086929	933658868	14-3-15 12:02:00	CHLORIDE	105.00	105
33086929	933658868	14-3-15 12:02:00	E-GFR, AA	9999999.00	> 60
33086929	933658868	14-3-15 12:02:00	E-GFR, NON-AA	9999999.00	> 60
33086929	933658868	14-3-15 12:02:00	HEMATOCRIT	41.80	41.8
33086929	933658868	14-3-15 12:02:00	HEMOGLOBIN	14.20	14.2
33086929	933658868	14-3-15 12:02:00	WHITE BLOOD CELL COUNT	5.60	5.6
33086929	933658868	14-3-15 12:02:00	UREA NITROGEN, BLOOD	15.00	15
33086929	933658868	14-3-15 12:02:00	GLUCOSE	93.00	93
33086929	933658868	14-3-15 12:02:00	CALCIUM	9.40	9.4
33086929	933658868	14-3-15 12:02:00	PLATELET COUNT	228.00	228
33086929	933658868	14-3-15 12:02:00	SODIUM	138.00	138

Table 3.3: The long-form lab data for patient PATNUM 33086929.

Table 3.3 displays the lab data extracted for patient 33086929. While there are only three different days for which lab data was collected for 33086929, many values were collected each day. Note that there are two version of the values recorded: `NUM_VALUE` and `ORD_VALUE`, which are almost identical except for when `ORD_VALUE` reports values such as > 60 (which only seems to happen for E-GFR measurements). The numeric version of these values is recorded as 9999999. Since this information is fairly useless, we used the `NUM_VALUE` column, but converted the 9999999 values to `NA` missing values.

PATNUM	DATE	CHLORIDE	GLUCOSE	HEMOGLOBIN	PLATELET COUNT	...
33086929	2015-03-14	105	93	14.2	228	...
33086929	2015-03-20	103	151	13.1	222	...
33086929	2015-03-21	100	124	13.7	228	...

Table 3.4: The wide-form lab data for patient PATNUM 33086929.

To convert each data source to a consistent format so that we can eventually join it to the denominator file, we focus on converting each dataset to a wide-format, where each row contains all of the measurements taken on a single day for an individual patient. This means that if there were multiple measurements taken for a lab on a single day, we summarized it by taking the average value for that day.

The corresponding wide-form data for patient 33086929 is presented in Table 3.4. Note that we no longer include the encounter ID `ENCNUM`, since we will be matching over patient ID, `PATNUM`. Since there were only three days represented in Table 3.3, the new wide-form version of the lab data consists of only three rows.

Figure 3.3 shows the distribution of each lab measurement across the entire dataset. Many of the labs are approximately symmetric (Hematocrit, Red Cell count, Hemoglobin, Chloride, etc), whereas some are heavily right-skewed (Alanine Transferase, Alkaline Phosphatase, Lymphocytes, Monocytes, etc).

Unfortunately, as is often the case with EHR databases, data coverage was not 100%. Many of the lab measurements were rarely recorded (presumably because these lab tests were not done). Figure 3.4 presents curves that display the proportion of patients with at least *one* measurement for each lab between the time on the x-axis and surgery (the right-most x-coordinate in the plot). Notice that none of the lab types were reported in more than 75% of the patients.

Since labs are typically taken in groups, where related labs are all measured at the same time, we see that there are similarly groups of labs that have similar reporting prevalence. Due to limited space on the plot, only one name from each group is reported next to overlapping curves. The most widely measured labs are blood counts including hematocrit, hemoglobin, platelet count, red cell count, and white blood cell count (all listed under hematocrit), and even these are only recorded for 75% of patients within the 30 days before surgery, and in less than 50% of patients in the week before surgery.

Metabolic labs including albumin, aspartate transaminase, alkaline phoshatase, and bilirubin are only measured in less than half of all patients in the 30 days before surgery. Other metabolic labs including C-reactive protein and creatinine serum are very rarely measured at all, with less than 10% of patients having a single measurement in the 30 days prior to surgery. We will discuss how we deal with missing data in Section 3.5.

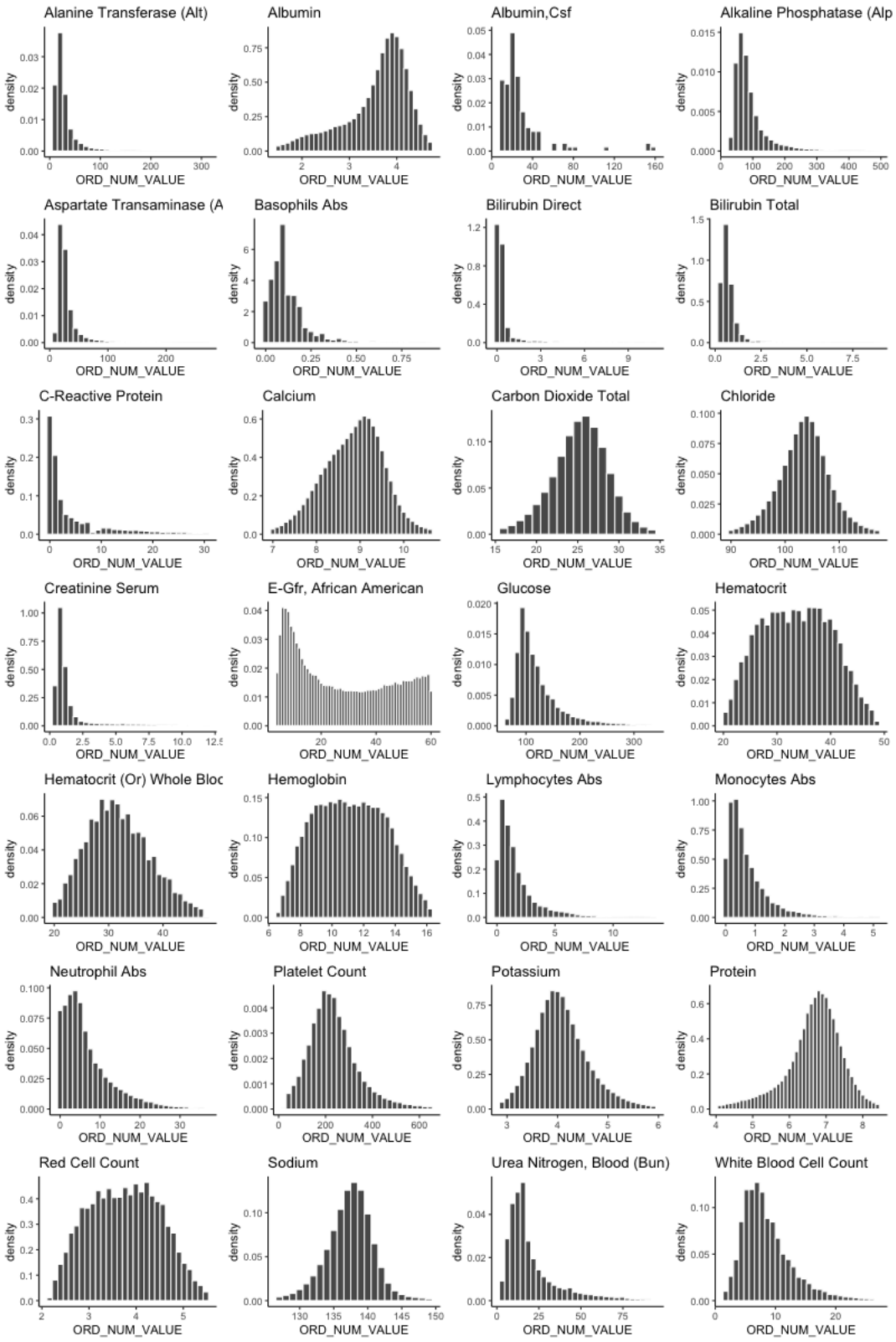

Figure 3.3: Histograms showing the distribution for each lab measurement across the entire dataset.

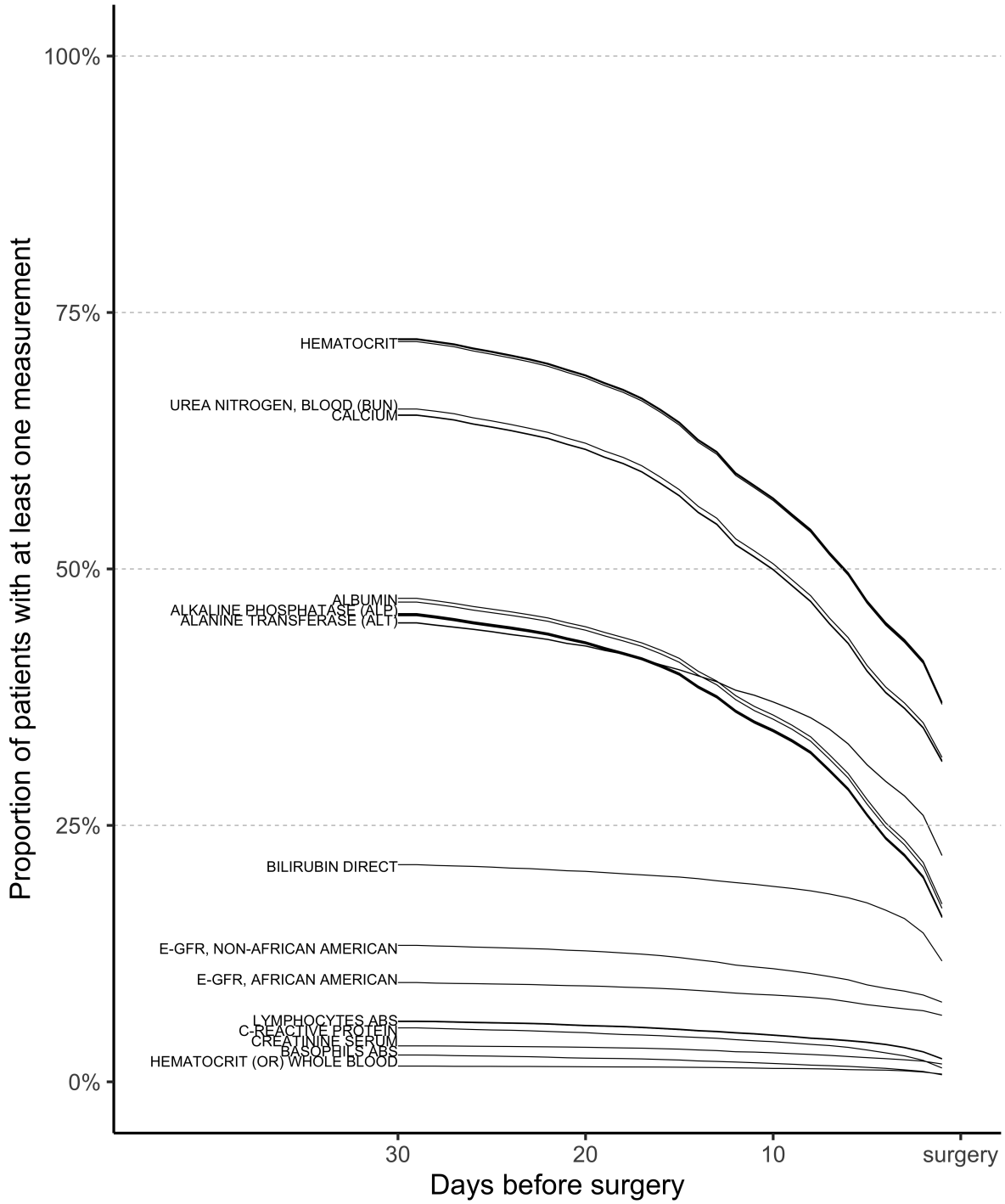

Figure 3.4: Line graphs displaying the proportion of patients with at least *one* measurement for the given lab between the time on the x-axis and surgery. Due to limited space on the plot, only one name from each group is reported next to overlapping curves.

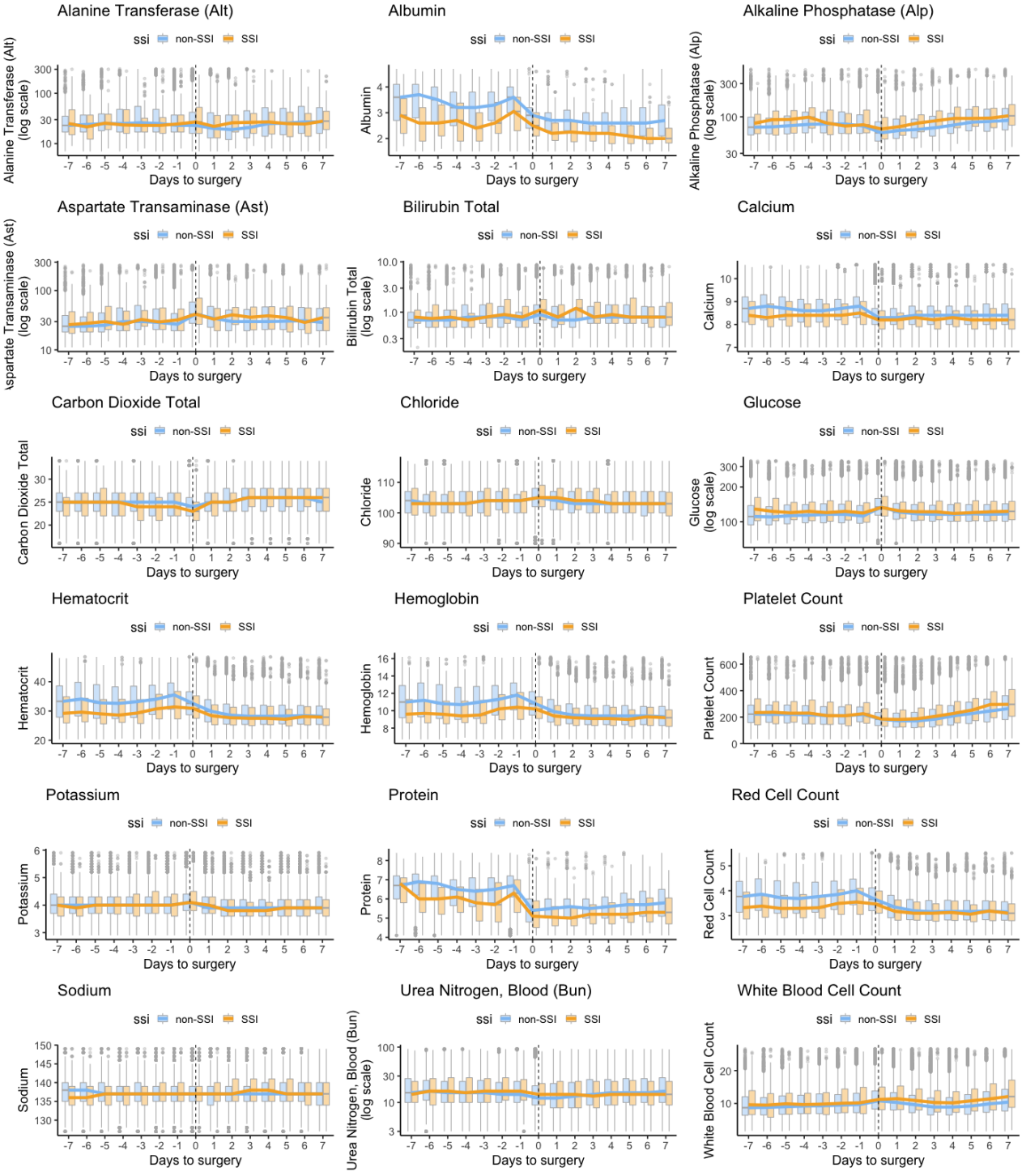

Figure 3.5: Boxplots and line graphs the distribution and median daily (relative to surgery) value of each lab for the SSI and non-SSI patients. The surgery takes place at time 0, and is represented by a vertical line.

Figure 3.5 displays boxplots and line graphs for the median daily value of each lab (where time is measured relative to surgery) for the SSI and non-SSI patients. We only include the labs that have at least one measurement for 40% of the patients 30 days before surgery. For the labs that were severely right-skewed, such as alanine transferase and bilirubin direct, we presented the y-axis on a log_{10} scale. Note that the orange (SSI) and blue (non-SSI) colors used in Figure 3.5 will be consistently used throughout the remainder of this thesis to refer to SSI and non-SSI patients, respectively.

From Figure 3.5, there are some clear differences between the SSI patients and the non-SSI patients for several labs, including albumin and the blood counts such as hematocrit, hematocrit, hemoglobin and red cell count. In each case, the non-SSI values are typically higher. The trends of the lab measurements before and after surgery are also interesting. For instance, albumin and protein levels drop dramatically after surgery, before increasing very quickly back to their original level. Hematocrit, hemoglobin, and red cell count levels also decrease after surgery, but do not recover their original values quite as quickly. Platelet counts also decrease, but quickly ascend to a level that is higher than the previous level (this makes sense since the patient's bodies will be producing platelets to heal the surgical wound).

PATNUM	ENCNUM	FLO_MEAS_NAME	MEAS_VALUE	DATE
33086929	915571140	PULSE	93.0	20-3-15 21:30:00
33086929	915571140	TEMP (CELSIUS)	37.5	20-3-15 22:00:00
33086929	915571140	R APACHE TEMPERATURE	37.5	20-3-15 22:00:00
33086929	915571140	PULSE	98.0	20-3-15 22:00:00
33086929	915571140	TEMPERATURE	99.5	20-3-15 22:00:00
33086929	915571140	PULSE	100.0	20-3-15 22:30:00
33086929	915571140	R APACHE TEMPERATURE	36.8	20-3-15 22:50:00
33086929	915571140	PULSE	98.0	20-3-15 22:50:00
33086929	915571140	TEMPERATURE	98.2	20-3-15 22:50:00
33086929	915571140	TEMP (CELSIUS)	36.8	20-3-15 22:50:00
33086929	915571140	TEMPERATURE	98.1	21-3-15 04:05:00
33086929	915571140	PULSE	81.0	21-3-15 04:05:00
33086929	915571140	TEMP (CELSIUS)	36.7	21-3-15 04:05:00
33086929	915571140	R APACHE TEMPERATURE	36.7	21-3-15 04:05:00
33086929	915571140	R APACHE TEMPERATURE	37.0	21-3-15 07:40:00
33086929	915571140	TEMP (CELSIUS)	37.0	21-3-15 07:40:00
33086929	915571140	PULSE	88.0	21-3-15 07:40:00
33086929	915571140	TEMPERATURE	98.6	21-3-15 07:40:00
33086929	915571140	TEMPERATURE	98.1	21-3-15 13:53:00
33086929	915571140	R APACHE TEMPERATURE	36.7	21-3-15 13:53:00
33086929	915571140	TEMP (CELSIUS)	36.7	21-3-15 13:53:00
33086929	915571140	PULSE	97.0	21-3-15 13:53:00
33086929	915571140	PULSE	92.0	21-3-15 19:30:00
33086929	915571140	R APACHE TEMPERATURE	36.8	21-3-15 19:30:00
33086929	915571140	TEMPERATURE	98.2	21-3-15 19:30:00
33086929	915571140	TEMP (CELSIUS)	36.8	21-3-15 19:30:00

Table 3.5: The long-form vitals data for patient 33086929.

Vitals EHR data

Much like the lab data, the vitals data was received in long-format. Table 3.5 shows the long-form data for patient 33086929. Notice that there are three separate temperature measurements taken: `TEMPERATURE`, `R APACHE TEMPERATURE`, and `TEMP (CELSIUS)`.

When creating a wide-form clean dataset, we only kept the temperature measurement that has the least amount of missingness across the data (the `TEMPERATURE` variable). We also group all observations made on the same day for the same patient together so that if multiple measurements were made on a single day, the average value was taken. The aggregated wide-form vitals data for patient 33086929 is shown in Table 3.6.

Approximately 85% of patients had at least one vitals measurement in the 30 days before surgery, and approximately 70% of patients had at least one measurement in the 7 days after surgery. Figure 3.6 displays the daily trends in pulse and temperature in the week before and after the surgery.

PATNUM	DATE	PULSE	TEMPERATURE	...
33086929	2015-03-20	93.6	98.4	...
33086929	2015-03-21	89.5	98.3	...

Table 3.6: The wide-form vitals data for patient 33086929.

For both the temperature and pulse measurements, there appears to be a small increase immediately following the surgery. The pulse measurements surrounding the surgery appear to be lower for the non-SSI patients than for the SSI patients.

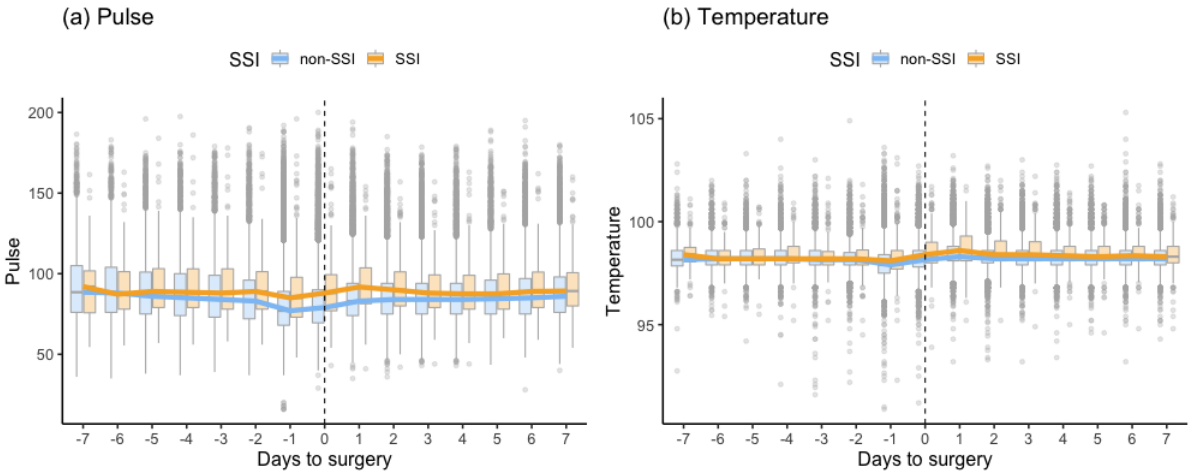

Figure 3.6: Boxplots displaying the distribution of daily (relative to surgery) (a) pulse and (b) temperature measurement for the SSI and non-SSI patients. The surgery takes place at time 0, and is represented by a vertical line.

Medication EHR data

The medication dataset contains all medications prescribed for the patient both historically and around the period of the surgery. Like the lab and vitals datasets, the medication dataset was also originally provided in a long format. However, instead of a medication name, the dataset referred only to a medication ID code and description (which included the name of the medication and the dosage). However, there were 22,449 unique medications in the original dataset, which is far too many to include as predictors in a predictive algorithm. Our collaborators were able to provide 50 groupings of the medication IDs, that we joined to the medication data as the `MEDICATION_CLASS` variable. The long-form data with the additional medication class variable for patient 33086929 is provided in Table 3.7. The full list of therapeutic classes can be found in the Appendix in Table A.4.

Similarly to the labs and vitals data, we created a wide-form version of the data with one row for the medications prescribed to a single patient on a single day. For the medications data, we created binary variables for whether each category of medications were prescribed on the given day. The wide-form data for patient 33086929 are shown in Table 3.8, indicating that they were prescribed anesthetics and antibiotics on March 20 and 21 2015, corresponding to the day of, and the day after their surgery.

Medication data within the 30 days prior to the surgery is available for 85% of patients.

PATNUM	ENCNUM	MED_ID	DATE	DESCRIPTION	MEDICATION_CLASS
33086929	941699452	5164	14-3-15	magnesium citrate oral solution	gastrointestinal
33086929	915571140	2000068	20-3-15	lactated ringers 1l	other
33086929	915571140	1800011	20-3-15	intraop NACL 0.9% 1000ml	other
33086929	930821965	4141	20-3-15	hydromorphone 1mg/ml	analgesics
33086929	930821965	139948	20-3-15	rocuronium 100 mg/10ml	autonomic_drugs
33086929	930821965	4896	20-3-15	lidocaine 20mg/ml	anesthetics
33086929	915571140	2706	20-3-15	dipenhydramine 50mg/ml	antihistamines
33086929	915571140	46610	20-3-15	paroxetine 30mg	psychotherapeutic_drugs
33086929	930821965	2000068	20-3-15	lactated ringers 1L	other
33086929	930821965	112220	20-3-15	ondansetron HCL 4mg/2ml	gastrointestinal
33086929	915571140	3327	20-3-15	fentanyl (PF) 50mcg/ml	analgesics
33086929	930821965	3846	20-3-15	glycopyrrolate 0.2mg/ml	gastrointestinal
33086929	930821965	142140	20-3-15	acetaminophen 1000mg/100ml	analgesics
33086929	915571140	109815	20-3-15	heparin, porcine 5,000unit/0.5ml	anticoagulants
33086929	915571140	1050137	20-3-15	intraop bupivacine 0.25%	other
33086929	930821965	6000	20-3-15	neostigmine methylsulfate 1mg/ml	autonomic_drugs
33086929	915571140	4141	20-3-15	hydromorphone 1mg/ml	analgesics
33086929	915571140	50039	20-3-15	zolpidem 5mg	sedative_hypnotics
33086929	915571140	1001234	20-3-15	cefazolin	antibiotics
33086929	915571140	6494	21-3-15	oxybutynin chloride 5mg	unclassified_drug_products
33086929	915571140	2241	21-3-15	docusate dosium 100mg	gastrointestinal
33086929	915571140	36448	21-3-15	hydrocodone 5mg	analgesics
33086929	915571140	43221	21-3-15	ciprofloxacin 500mg	antibiotics
33086929	915571140	148665	21-3-15	lidocaine 5%	anesthetics
33086929	901090256	11333	26-1-15	lisinopril 10mg	cardiovascular

Table 3.7: The long-form medication data for patient 33086929.

PATNUM	DATE	anesthetics	antibiotics	antidotes	antivirals	...
33086929	2015-01-26	0	0	0	0	...
33086929	2015-03-14	0	0	0	0	...
33086929	2015-03-20	1	1	0	0	...
33086929	2015-03-21	1	1	0	0	...

Table 3.8: The wide-form medication data for patient 33086929.

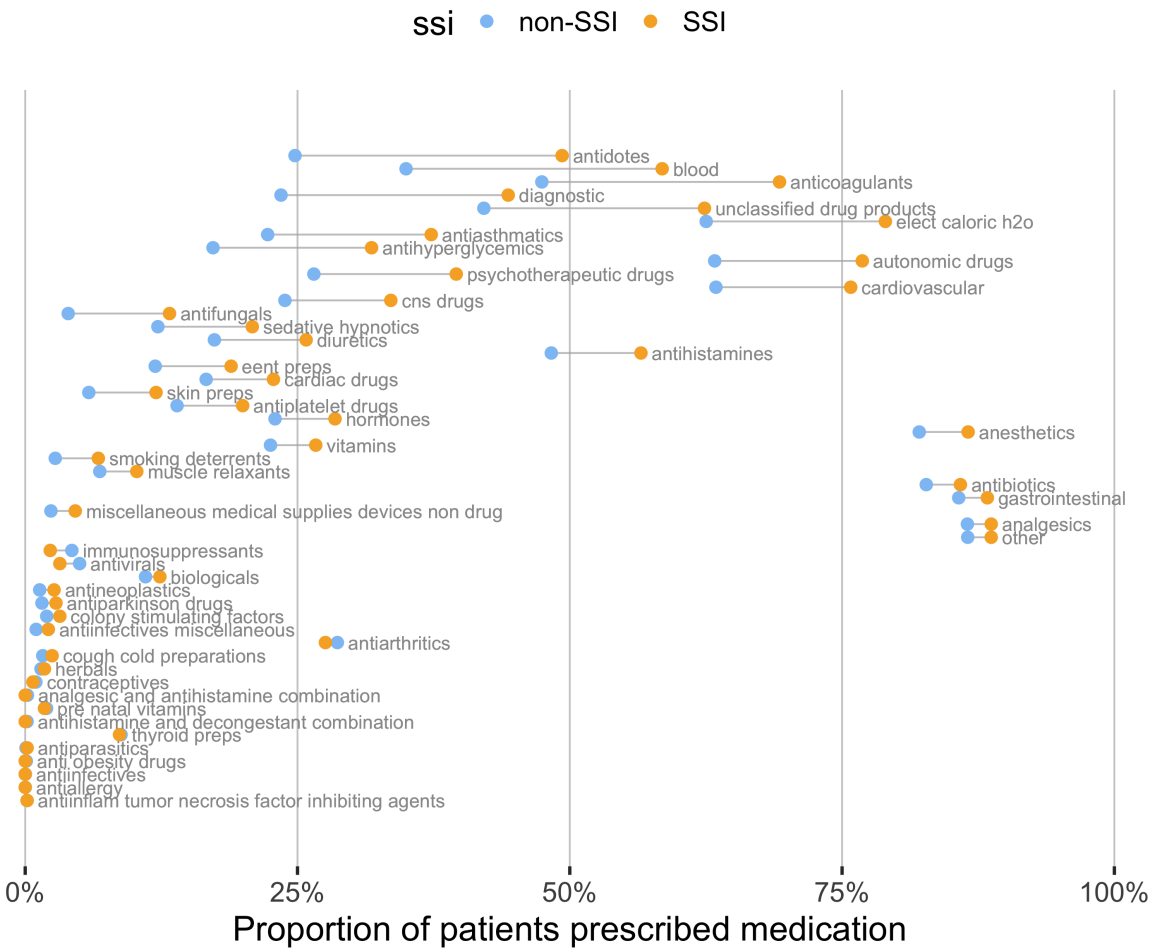

Figure 3.7: A dot plot displaying the different between the proportion of SSI and non-SSI patients prescribed each medication class. The y-coordinate corresponds to the medication classes arranged from top to bottom in decreasing order of difference between SSI and non-SSI prescription rates, and the x-coordinate corresponds to the SSI (triangle) and non-SSI (circle) prescription proportions. The line connecting the circle and triangle correspond to the difference in proportions of patients prescribed the medication.

Figure 3.7 explores the medication classes whose prescription rates differ most between the SSI and non-SSI patients. Prescriptions of antidoes (medication to counteract a poison), drugs related to blood (such as those prescribed for clotting and blood flow disorders), and anticoagulants (aimed at preventing the formation of blood clots) all differ the most between the SSI and non-SSI populations. Blood medications and anticoagulants are both related to how a wound heals, so this finding seems fairly intuitive.

Summary of EDA

In this section we undertook a thorough exploratory data analysis. We demonstrated the extent of missing values in the lab data, which will lead us to exclude many rarely measured lab values (such as C-Reactive Protein, and E-GFR). We showed that the vitals data did not suffer from the same extent of missing values (since over 70% of all patients had at least one temperature and pulse measurement in the month before surgery).

We also found that some of the features that appear to most distinguish between SSI and non-SSI patients include the length of surgery, whether or not the patient underwent a transfusion, several lab measurements (including albumin, hematocrit and red blood cell counts), pulse, and temperature measurements, as well as some medication classes such as antidotes, medications related to blood (e.g. for clotting disorders), and anticoagulants (aimed at preventing the formation of blood clots). We will see that many of these features are also identified by our algorithmic analysis in Chapter 4.

In the next section, we will discuss the pre-processing steps we take to deal with the missing data, categorical variables, and combining each of these sources of data into a single covariate matrix.

3.5 Data pre-processing

After each separate dataset had been cleaned, they were joined together to form a full covariate matrix. In this section, we describe our patient exclusion criteria, and the methodology for joining the datasets together, for splitting the data into training and testing datasets, handling missing values (including removing variables with large amounts of missingness, and imputing missing values), and converting categorical variables to dummy variables.

Exclusion criteria

Based primarily on discussions with our surgeon collaborators, we decided to exclude

- 60 patients with an infection already present at the time of surgery.
- 4,225 patients with missing time of surgery. While we could probably impute these values, our collaborators believe that these patients' data was collected under an old EPIC medical record system, which may have also had other data inconsistencies.

Indeed, the new EPIC system was introduced in July 2014, and 95% of the patients with missing surgery times had their surgeries in early 2014. Figure 3.8 displays a histogram of the surgery dates for the patients with missing surgery times.

- 2,926 children under the age of 18.

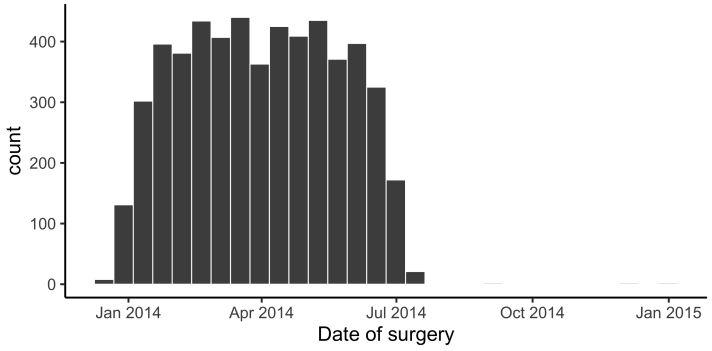

Figure 3.8: A histogram showing the surgery dates for patients with missing surgery times.

Creating the covariate matrix: combining sources of data

We joined the different data sources together using PATNUM as a key. However, since we want our final covariate matrix to contain one row per procedure, and our lab, vitals, and medication data each have one row per day on which a measurement was taken, we need to further aggregate these datasets so that we have aggregate these rows together. We do this by taking the maximum and minimum vitals and lab values over the 30 day pre-surgery period, as well as the maximum and minimum vitals and lab values over the post-surgery period up to the time the model is being implemented (which in this thesis is 7 days). Similarly, we aggregate the medication data to only include those medications prescribed in the 30 days prior to and the 7 days after surgery.

While we ended up choosing the maximum and the minimum, we considered various statistics including the maximum, median, minimum, mean, standard deviation, and range. Since many of the patients had only one value in the 30-day period pre-surgery, the range and standard deviation were not useful measures. We found that the maximum and minimum were both highly correlated with the mean and median, but not so correlated with one another. Our decision to thus focus on the maximum and minimum values was based on these findings and discussion with our collaborators who agreed that the maximum and minimum measurements were a reasonable summary. We later tried including the mean and median and found no noticeable difference in our downstream results.

After joining together the NHSN denominator with the aggregated lab, vitals, and medications data, our covariate matrix has 263 variables and 37,881 rows. These variables consist of

- 7 patient variables from the NHSN denominator file.
- 18 surgery variables from the NHSN denominator file.
- 120 lab variables (each combination of pre-surgery/post-surgery and minimum/maximum values for 30 different lab measurements).
- 8 vitals variables (maximum and minimum temperature and pulse both pre-surgery and post-surgery)
- 80 medication variables (37 post-surgery and 43 pre-surgery).
- 1 SSI variable.

We also computed several additional variables from the data including

- The length of the surgery (the difffernce between the in-time and the out-time).
- The difference between the average temperature and pulse in the 30 days before and after the surgery.
- A risk category (low, medium, high) for each surgery identified by our surgeon collaborators at UC Davis

The risk categories for each procedure are listed in Appendix Table A.1.

Splitting into training and testing sets

To compare different modeling approaches on independent testing data, we split the data into a training set (60%) and a test set (40%). We did not include a validation set in our split for two reasons: (1) in future work to be completed after the publication of this thesis, we will be validating the model on an independent dataset from the Veterans Affairs (VA) hospital in Davis, and (2) the number of SSI cases in the test data would have been too small to do appropriate validation of the model with only 20% test/validation samples. To ensure that the test set appropriately represents the individual procedures and SSI patients, the train-test split was conditional on procedure and SSI status.

Removing variables with too many missing values

First, we removed any variable that had more than 30% of its values missing (so the variable needed to have at least 70% non-missing values). This cutoff was based both on a judgement call and a natural break in the distribution of missing values among the variables shown in Figure 3.9. Almost all of the variables removed using this cutoff are lab variables, with two exceptions: surgeon age and surgeon physician type, which are both surgery variables from the NHSN demoninator dataset.

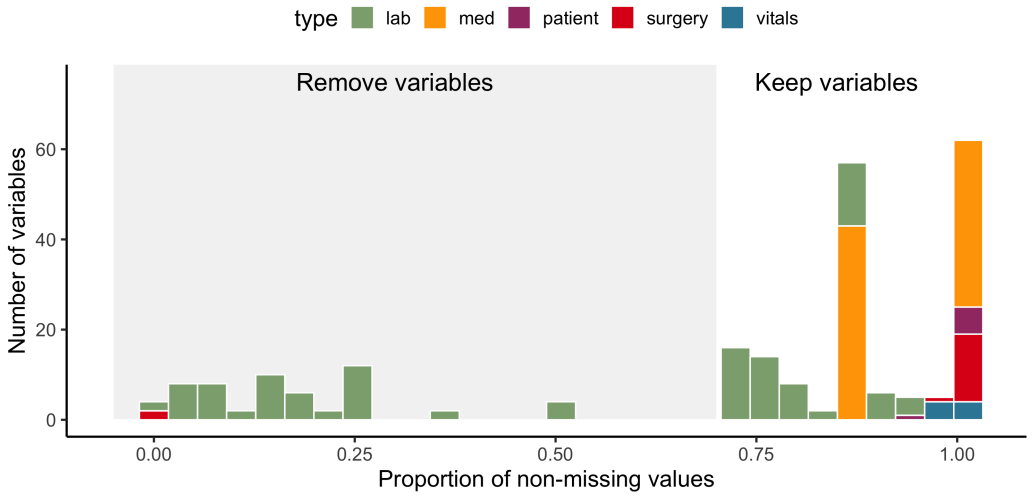

Figure 3.9: A histogram displaying the proportion of missing values across the variables in the covariate matrix. The grey area corresponds to the variables that fall below the 70% non-missing threshold that will be removed.

Imputing missing values

To impute the remaining missing values, we used a method called missForest, which uses the Random Forest (RF) algorithm which predicts missing values using a RF trained on the observed parts of the dataset[92, 13]. We implement missForest using the the *missRanger* R package. We were careful not to use post-surgery variables to impute pre-surgery variables.

To examine whether the imputed variables were different to the original variables, we implemented permutation tests based on 1000 permutations for each variable. These tests were implemented using the `coin` R package [40]. While many of the p-values were significant, the actual raw mean value differences were so small that we weren't too concerned. For instance, the mean un-imputed BMI was 28.95 (SD 8.25), while the mean imputed BMI 28.67 (SD 8.21), but this difference was statistically significant due to the large sample size.

Boxplots displaying the imputed and unimputed distributions of a random selection lab variables (each with less than 30% missingness) are shown in Figure 3.10. The inter-quartile range of the imputed variables seems to be slightly narrower than the original unimputed variables (this is particularly noticeable for the carbon dioxide, chloride, and sodium lab variables). We compare our final results with this missForest-based imputation with the results that we obtain based on mean imputation, and don't notice a major difference, primarily because the majority of the lab variables don't end up being extremely important in our predictions.

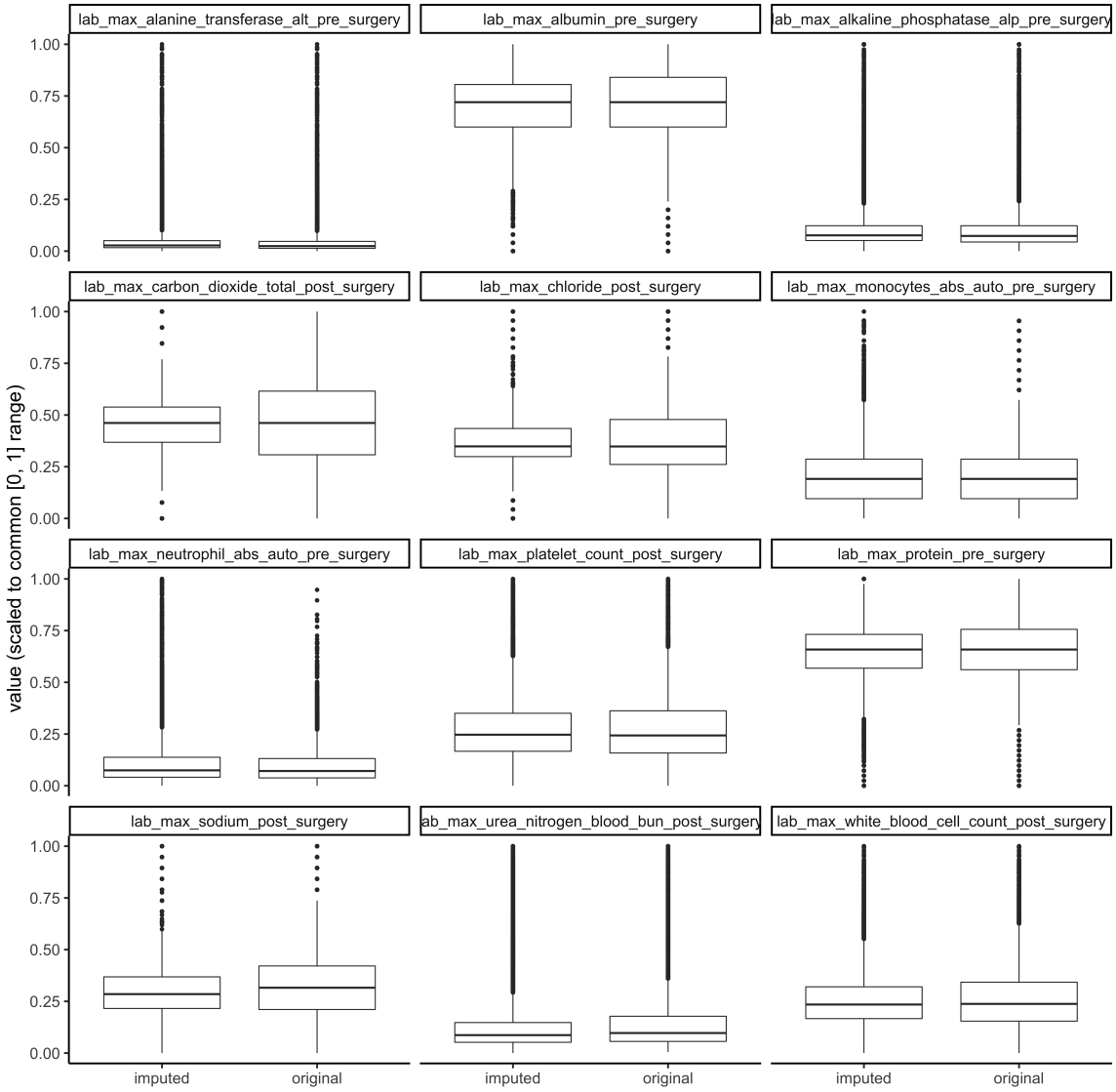

Figure 3.10: Boxplots comparing the distribution of the original observed values and the imputed values for 12 randomly selected lab variables.

Converting categorical variables to binary dummy variables

After imputing our data, we converted each of the categorical variables with more than two categories to binary dummy variables (with a reference value removed). For instance, we converted the primary surgery diagnosis variable into 19 binary variables, one for each of the 20 diagnoses (minus one diagnosis which acts as the reference class).

3.6 Conclusions

While SSI is relatively uncommon (occurring in only 2-5% of surgeries), it can lead to critical consequences for patient outcomes and hospital finances. In the 1990s, the CDC began implementing mandatory surveillance screening procedures, leading to the development of the NHSN database to which hospitals are required to report incidences of SSI.

While these NHSN databases alone have been used to develop predictive models, very few works have tried to combine these databases with readily available EHR data (including labs, medications, vitals, and diagnoses) that exists in all hospitals to develop more accurate predictive models.

In this chapter, we developed a combined covariate matrix consisting of NHSN data and EHR data both from UC Davis. We catalogued the data cleaning and pre-processing steps we implemented to create this covariate matrix, and we provided some preliminary exploratory analyses outlining how individual variables from each dataset might be associated with SSI.

Chapter 4

Predicting Surgical Site Infections

In the previous chapter, we outlined the need for an accurate predictive model for Surgical Site Infections (SSI) and discussed the potential benefits that could arise from combining the NHSN or NSQIP SSI surveillance data with routinely collected variables from Electronic Health Records (EHRs), specifically lab, vitals, and medication data.

In this chapter, we implement and evaluate a predictive modeling framework using the combined UC Davis NHSN and EHR datasets introduced and explored in the previous chapter. The modeling framework we will introduce is based on Random Forest (RF) and is adapted to the rare-event scenario using repeated random subsampling to create many balanced subsamples. While this technique is not widely used for dealing with class imbalance, it has been described in a few influential works including [13, 80], and we will show that it leads to better predictive performance than traditional approaches dealing with class imbalance such as upsampling and downsampling.

Our procedure involves fitting a RF model to each balanced subsample and extracting the predicted "probability" of SSI, which we then aggregate across the forests by taking the average "probability". Since this mean value is not quite a probability in the traditional sense, we will call this averaged value the "SSI score". The UC Davis cohort of patients has been split into a 60% training set and 40% testing set, and the model will later (i.e. after publication of this thesis) be evaluated on an independent cohort of patients from the Davis VA hospital (in an example of transfer learning [77]). In this thesis, we focus on predicting SSI status (i.e. whether the patient will develop an SSI within 30 days of the surgery) using data up to 7 days post-surgery.

Section 4.1 describes the current approaches to predicting SSI in the literature.

4.1 Existing approaches to predicting SSI

The availability of surveillance data, such as the NHSN and NSQIP databases, collected on SSI patients (as well as non-SSI control patients) has spurred several efforts for developing

generalizable predictive modeling methods to identify patients who are at increased risk for SSI.

In the 1990s, the NHSN itself developed the NNID Risk Index model that consists of 3 binary variables: ASA score (3, 4, or 5), wound classification (contaminated or clean), and procedure duration in minutes (whether the surgery length was greater than the 75th percentile or not). Each risk factor represents 1 point, so the NHSN SSI risk index ranges from 0 (lowest risk) to 3 (highest risk) [38, 23]. This model was extended in 2011 by researchers from the CDC in [71], where they used "stepwise-logistic regression" models to develop risk models by procedure categories. They used the original three variables as well as additional variables of convenience that are routinely reported to the NHSN as part of the existing SSI surveillance methodology including general anesthesia, emergence procedure, gender, trauma association, medical school affiliation, number of hospital beds, and age. Since the researchers worked for the CDC, they had access to a vast database consisting of the NHSN data from 847 hospitals with a total of 849,659 procedures. They fit models separately for each procedure, and found that their AUC values ranged from 0.59 to 0.85, which was slightly higher than the original NHSN risk index models. This is the model used to predict infection risk by the NHSN today.

Beyond these efforts by the CDC, there are surprisingly few strains of research focused on developing predictive models for SSI. One example includes [104], who fit models to NSQIP data, and grouped surgeries together based on the first three digits of their CPT (Current Procedural Terminology) code. They developed a "CPT3 score" for each group of procedures (based on the first 3 digits of the CPT code) that corresponded to the ratio of observed to expected SSI cases in the procedure group. Based on training and validation sets each of approximately 180,000 patients, their logistic regression model that included this CPT3 score as a variable reported a c-statistic of 0.8. They also developed an SSI risk scoring procedure that can be computed without a computer based on wound type, outpatient/inpatient, ASA class, BMI, surgery length, peripheral vascular disease, septic, wound type, and their CPT3 score. There also exist some predictive methods that focus on specific subdomains of organ system procedures, such as colon cancer, cardiac, neurological and others [53, 49]. However, these results do not extend beyond the NSQIP database.

Combining NHSN or NSQIP data with EHR data

Despite the increasing availability of other sources of data, we could only identify one study in the literature that attempts to combine data from the EHR (such as lab, vitals, diagnosis, and medication data) with the data collected from the NHSN or NSQIP databases. [91] developed methodology for predicting SSI by combining the typical patient and procedural information with comorbidities and lab results. However, rather than building and applying their Support Vector Machine (SVM)-based methodology to a real cohort of patients, they curate a small *matched* cohort of 1,000 patients undergoing gastrointestinal surgery, where 10% of the patients developed SSI and 900 did not. This SSI rate is more than double what is typically observed in reality. The authors themselves state that "'*the problem does not*

entirely reflect the clinical scenario". Thus while this was the only work we could find that included EHR data in their predictor set, the results are not applicable to a general setting.

Our approach in this thesis is to develop Random Forest (RF)-based models based on NHSN data as well as incorporating additional routinely collected EHR data including laboratory test results, such as blood counts and metabolic panels; vitals measurements, such as temperature and pulse; and medications prescribed, including antibitiotics and immunosuppressants. However, unlike [91], we will be applying our methodology to a real cohort of the approximately 30,000 patients who underwent surgeries at UC Davis between 2014 and 2018, and will develop a ready-to-use method that can be used by clinicians for this population.

4.2 Generating repeated balanced subsamples

Since we are dealing with a severely unbalanced dataset (only 2.5% of patients have an SSI), in order to best capture the data patterns that differ between the SSI and non-SSI classes, we generate many balanced subsamples of the data, where a balanced subsample is one that has the same number of SSI and non-SSI patients. The motivation behind using many balanced subsamples rather than a single downsampled dataset (where we take a subsample of non-SSI patients equal in size to the number of SSI patients) is that if we only ever look at a single set of a few hundred non-SSI patients, we are ignoring over 95% of our data. Conversely, if we use a single upsampled dataset (where we repeatedly sample the SSI patients until we have a sample equal in size to the non-SSI patients), each SSI patient would appear in the data approximately 40 times, leading to disproportionate influence of individual SSI patients.

Similar approaches to our repeated balanced subsampling procedure have seen success in the literature [60, 20], and we show in Section 4.7 that the repeated balanced subsampling approach performs better than traditional upsampling and downsampling approaches.

To generate a balanced subsample, the majority class (non-SSI) is extremely downsampled and the minority class (SSI) is slightly downsampled so that each model is trained on an equal number of SSI and non-SSI patients (equivalent to 70% of the SSI patients; around 460 patients in each class). So that our results are not strongly influenced by our random sampling, we repeat this procedure 1000 times [18]. Figure 4.1 depicts the generation of three balanced subsamples, to which Random Forest models are then fit.

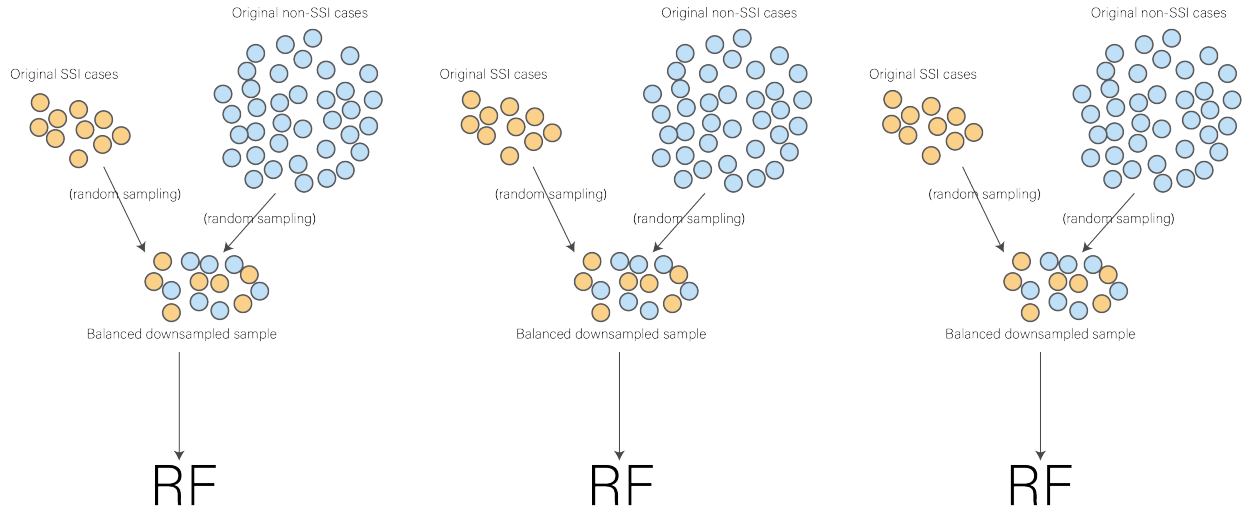

Figure 4.1: The downsampling procedure: a 70% random subsample of the minority SSI class is taken, and an equal sized subsample is taken from the majority non-SSI class.

4.3 Feature selection

After implementing the pre-processing steps outlined in Section 3.5, we have 239 variables. Before pre-processing, the data consisted of 263 variables. During pre-processing we lost many variables due to missing values, but also gained many variables by splitting categorical variables into dummy variables and taking various summaries of lab and vitals measurements.

Since we have only around 600 SSI patients in our training data, but tens of thousands of non-SSI patients, we perform some preliminary feature selection before we begin modeling in order to offer the best chance of capturing meaningful patterns in the data and to avoid overfitting.

Feature selection is implemented by fitting Random Forest models to the balanced subsamples, not with the aim of generating a prediction, but with the aim of extracting variable importance scores. An RF model is fit to each balanced subsamples using the *ranger* R package [112], and the gini impurity variable importance scores are extracted from each model. We examine the importance scores across 1000 models so that we are capturing a wide range SSI and non-SSI samples, and so that we can assess the stability of the variable importance for each feature across different subsets of the data. For comparison, we also fit a single RF model using the entire unbalanced dataset, and found that the top 50 variables from the single unbalanced model were very similar, but the order was somewhat different.

Figure 4.2 presents boxplots displaying the distributions of the 1000 importance scores for top 50 variables. Notice that the variables with higher importance scores also tend to have higher variability across the resampled models.

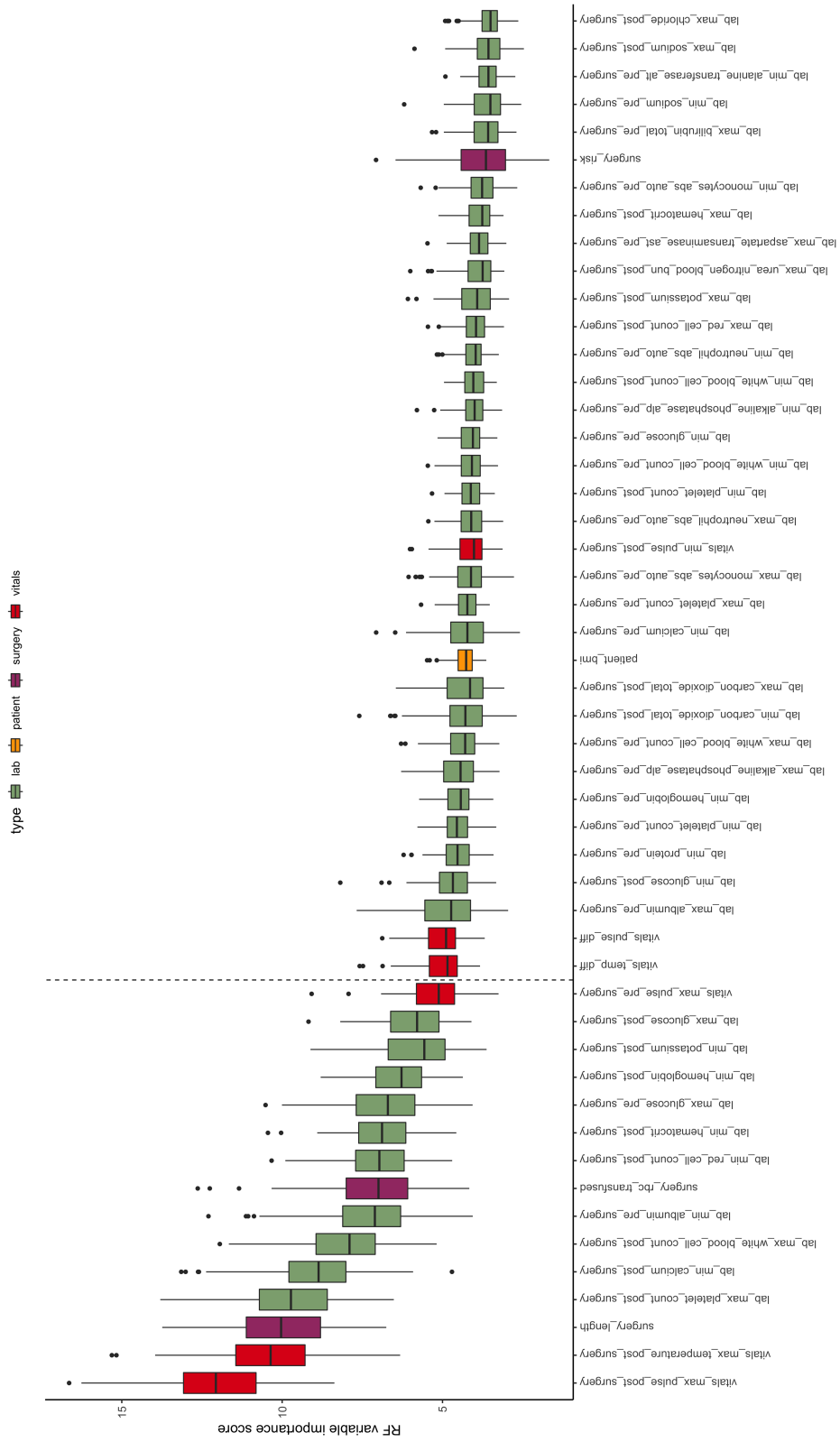

Figure 4.2: Boxplots displaying the distribution of importance scores across the bootstrapped downsampled balanced RF models. The vertical line represents the top 15 features.

The 15 most important variables based on the resampled importance scores are

1. maximum pulse following surgery
2. maximum temperature following surgery
3. the length of the surgery
4. maximum platelet count following surgery
5. minimum calcium value following surgery
6. maximum white blood cell count following surgery
7. minimum albumin level prior to surgery
8. amount of red blood cells transfused
9. minimum red blood cell count following surgery
10. minimum hematocrit measurement following surgery
11. maximum glucose level prior to surgery
12. minimum hemoglobin measurement following surgery
13. minimum potassium measurement following surgery
14. maximum glucose level following surgery
15. maximum pulse prior to surgery

When shown to our medical collaborators at UC Davis, they described realistic domain reasons for each of these variables being related to SSI. A future experiment to strengthen this finding will involve showing a group of medical professionals two lists of variables, such as this list plus another list (such as the next 15 most important features), and have them decide which list contains more important variables for predicting SSI.

4.4 The SSI model

The model-fitting procedure is very similar to the feature selection procedure that we just introduced in Section 4.3. However, instead of fitting RF models using all of the features, we only use the top 15 important features described above. Specifically, we fit 5,000 RF models each based on the top 15 variables identified in our feature selection analysis above.

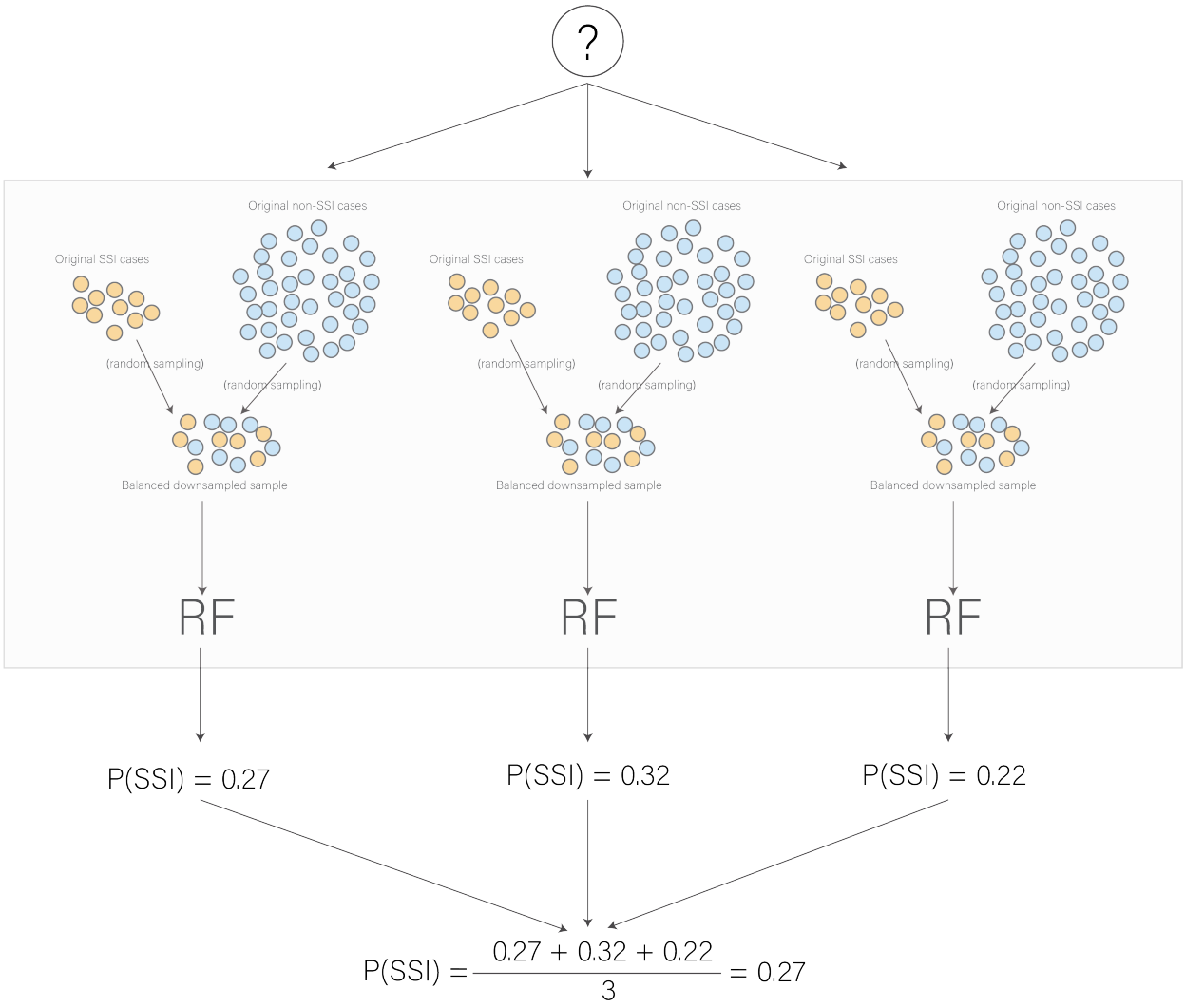

Figure 4.3: The prediction procedure for a new patient.

Each RF model is based on a balanced sample of 460 randomly selected SSI patients (corresponding to a 70% subsample of the SSI class), and an equal number of 460 randomly selected non-SSI patients. The predicted response for a new patient is the average predicted "probability" (across all of the 5,000 forests) that the patient is in the SSI class as outlined in Figure 4.3. We call this average predicted "probability" the *SSI score*.

4.5 SSI score performance evaluation

First we assess the performance of the aggregated models on the training data only. Keep in mind that many of the non-SSI training cases will not have appeared in the subsamples used to build the RF models. Figure 4.4 displays the density of the SSI score for the SSI

and non-SSI patients. There is a very nice separation between the two densities, with the SSI score for non-SSI patients being right-skewed (concentrated at lower values), and the SSI score for SSI patients being left-skewed (concentrated at higher values). The AUC for the training set is extremely high, at 0.976.

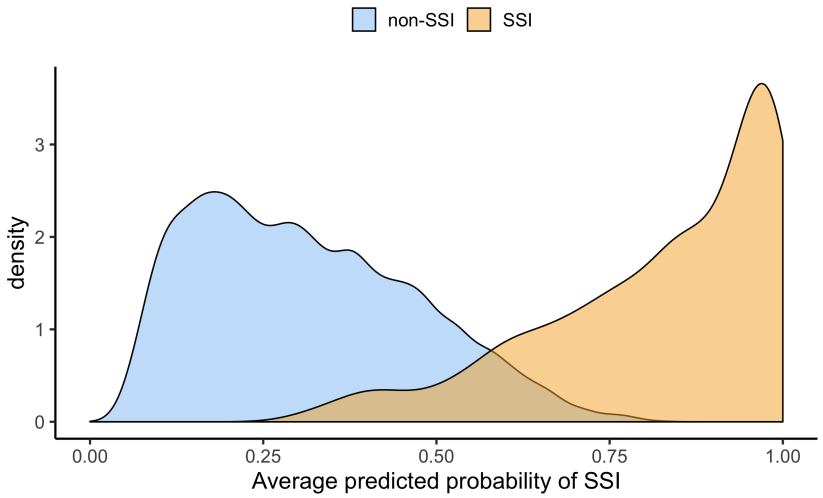

Figure 4.4: A density plot comparing the distribution of the average predicted SSI probability (SSI score) across the 1000 RF models for the SSI and non-SSI training patients.

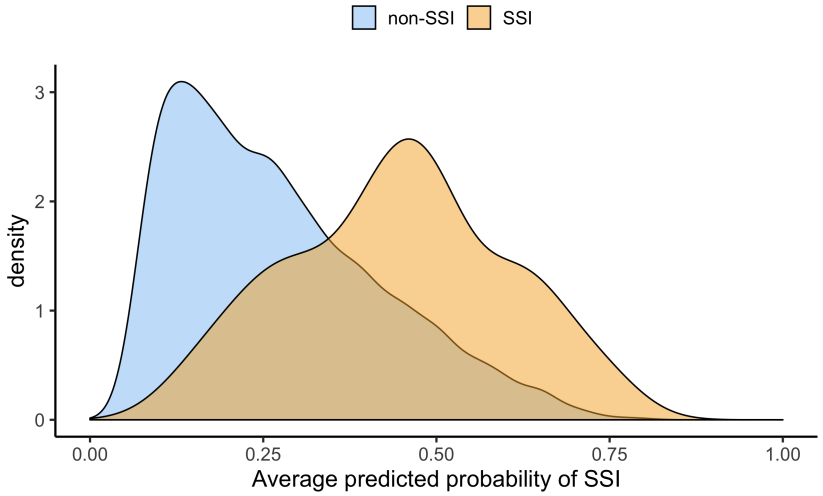

Figure 4.5: A density plot comparing the distribution of the average predicted SSI probability (SSI score) across the 1000 RF models for the SSI and non-SSI test set patients.

Naturally, the next question is whether this performance also extends to the UC Davis test set patients who were not used in the model training. Figure 4.5 displays the same densities as Figure 4.4, but for the test set patients instead of the training set patients. There is still a clear distinction between the densities, however the probabilities themselves have all shifted to the left (lower values) for both classes, but especially for the SSI patients.

There is unsurprisingly more overlap between the SSI and non-SSI densities for the test set, but there is still a clear separation, and overall the test set performance is still quite good. An ROC curve for the test set patients is displayed in Figure 4.6, and we show that the performance remains nearly identical (but slightly worse) when we include 25 features (rather than 15), and although we don't display it here, this remains true when we include even more features and when we include fewer features. The test-set AUC for the model built on the top 15 features is 0.79.

Once we have access to the data, we will also validate these results on validation data from the VA hospital.

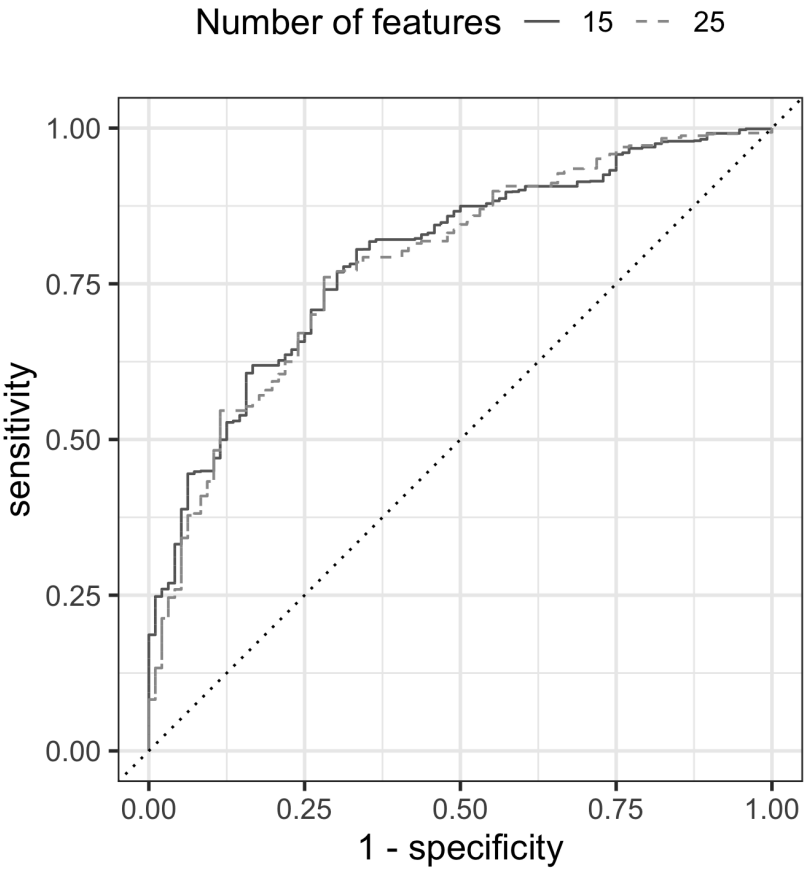

Figure 4.6: A test-set ROC curve for the model built on 15 features (solid line) and the model built on 25 features (dashed line). The AUC for the model built on 15 features 0.792, and the AUC for the model built on 25 features is only slightly lower at 0.783.

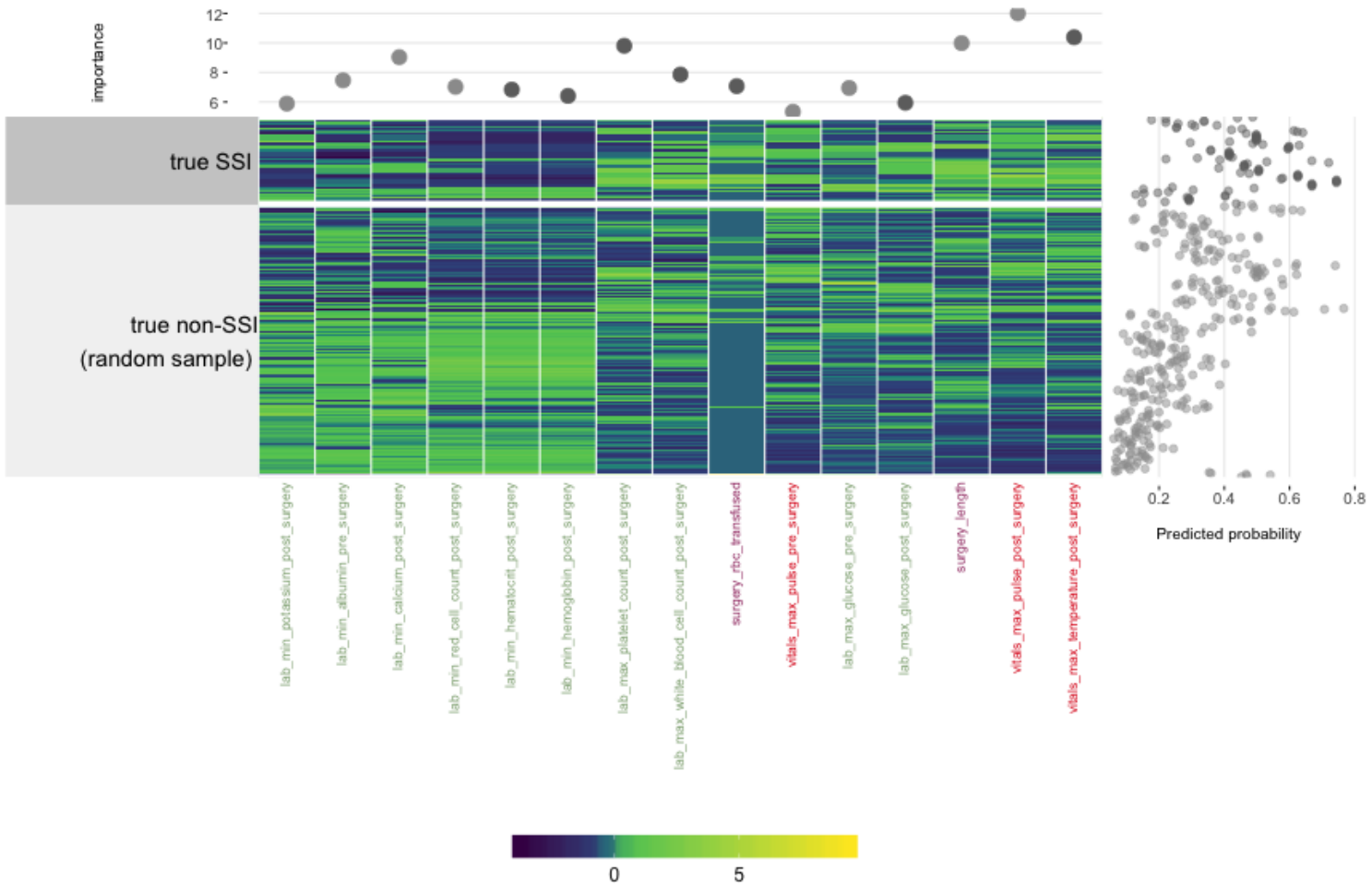

Figure 4.7: A superheatmap displaying the values of each variable for the 96 test-set SSI patients and a random sample of 300 of the 7,917 non-SSI test-set patients. The variable names are colored by type. Hierarchical clustering is used to arrange the rows and columns. The variable importance is plotted above the heatmap, and the SSI score is plotted to the right of the plot.

In Figure 4.7, to obtain a clearer picture of how the model is performing on the individual patients, we use the superheat R package that we introduced in Part 1 of this thesis [8]. Due to the gross imbalance of SSI to non-SSI cases, instead of visualizing the entire test set, we instead visualize all 96 SSI test set cases, but only a random sample of 300 of the 7,917 non-SSI test set cases. The rows (patients) are arranged using hierarchical clustering, so that similar patients appear next to one another. The SSI score is plotted as a scatterplot to the right of the heatmap. We see that there are a number of non-SSI cases (the group of rows at the top of the non-SSI block) whose lab data looks a lot like the SSI cases, and these patients also have a higher SSI scores. Similar patterns are seen with different random samples of the non-SSI test cases.

Figure 4.8 displays the proportion of test set patients that have an SSI for each SSI score rounded to the nearest decimal point. As the SSI score increases, so too does the rate of SSI. As the SSI score becomes greater than 0.5, the rate of SSI in the test set patients surpasses the average SSI rate across the entire training population.

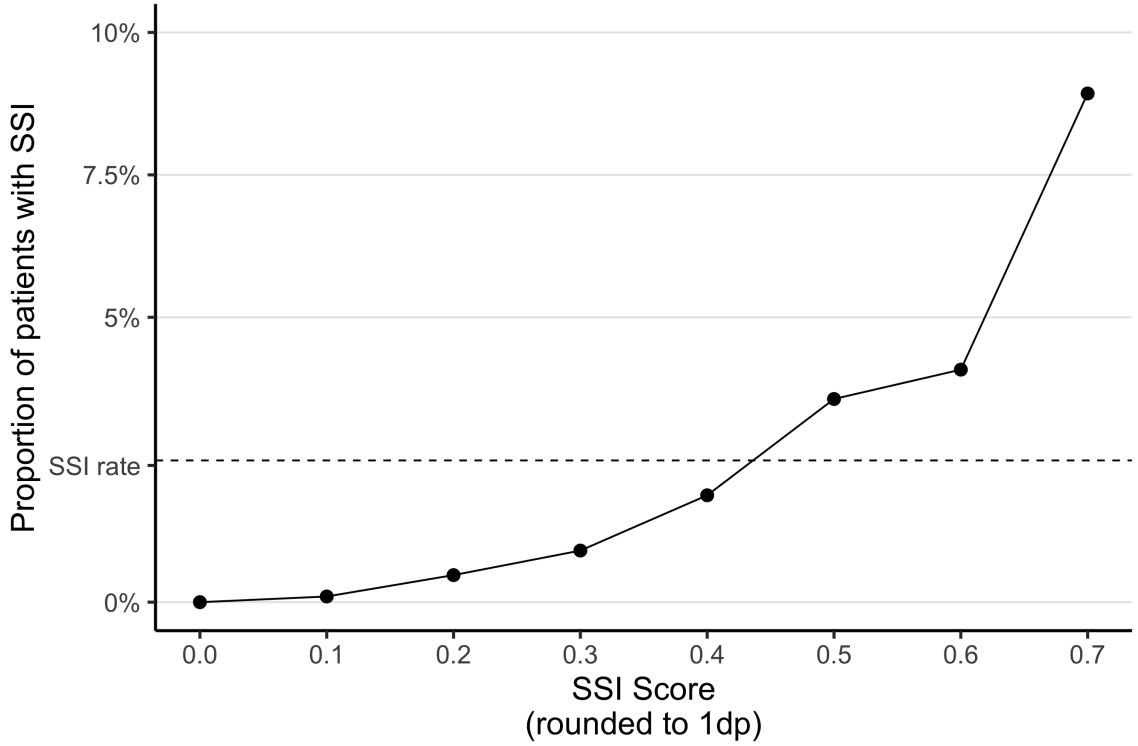

Figure 4.8: A line plot displaying the SSI rate among the test set patients that attained each SSI score rounded to the nearest decimal point. The horizontal dotted line corresponds to the overall proportion of SSI in the training data.

4.6 Identifying interactions

To examine whether there are any meaningful feature interactions driving the predictive performance of the RF models, we use the iterative Random Forest (iRF) method developed by [9, 55]. iRF identifies feature interactions by identifying groups of features that fall along the same decision paths across many bootstrapped versions of RF models.

Since the prediction problem itself is very difficult due to the extreme class imbalance, identifying interactions between features also proved very difficult (even when reducing the dimensionality using techniques such as PCA, ICA and supervised PCA). However, in a vein similar to our balanced subsample approach to prediction, we identified interactions on 10 different randomly selected balanced subsamples, and found that there were a few interactions that arose across almost all of the balanced subsamples.

Specifically, we filtered to interactions that were shown to

- improve the predictive power when compared to just the individual features alone (mean improvement in precision; MIP) in at least 80% of internal RF bootstrap replicates
- that were found along at least 10% of decision paths leading to an SSI prediction (prevalence) - this indicates that the interaction is sufficiently widespread
- for which at least 60% of the SSI patients fall into leaf nodes that contain the interaction (precision) - this indicates that the interaction is enriched for SSI patients relative to non-SSI patients
- that were found in at least 80% of internal RF bootstrap replicates (stability).

We found that the following four interactions arose from 9 of our 10 balanced subsamples:

- high platelet count and high pulse after surgery (mean precision: 0.74; mean prevalence: 0.19, mean MIP: 0.06; mean stability 0.98)
- long surgery length and high pulse after surgery (mean precision: 0.73; mean prevalence: 0.28, mean MIP: 0.07; mean stability 1.00)
- long surgery and high temperature after surgery (mean precision: 0.74; mean prevalence: 0.22, mean MIP: 0.09; mean stability 1.00)
- high pulse and high temperature after surgery (mean precision: 0.72; mean prevalence: 0.30, mean MIP: 0.07; mean stability 1.00)

and the following two arose from 7 of our 10 balanced subsamples.

- long surgery and high platelet count after surgery (mean precision: 0.76; mean prevalence: 0.19, mean MIP: 0.07; mean stability 1.00)

- high platelet count and high temperature after surgery (mean precision: 0.76; mean prevalence: 0.17, mean MIP: 0.07; mean stability 0.97)

In the end, none of these interactions are particularly surprising. If the surgery is longer, then there is more opportunity for infection, which will typically be manifested as a fever (high temperature and high pulse) and high platelet count after surgery (often indicative of inflammation or infection). While it is comforting to know that these intuitive interactions are identified in the data and by our modeling approach, they are unlikely to provide any new information to medical practitioners.

4.7 Comparing modeling approaches

In this section we explore the stability of our results to different analytic decisions [115]. For instance, we compare the performance if we had fit a single (i.e. not subsampled) model, or if we had used simple downsampling or upsampling procedures to balance the data, if we had fit separate models to each procedure risk group, if we had just used the NHSN variables (but not the EHR variables), or if we had used SVM or logistic regression classifiers.

Single model fit to full unbalanced dataset

In this section, we compare the subsampled balanced approach to modeling with the performance of a single model fit to the full unbalanced dataset. While the AUC only decreases slightly to 0.77, the bulk of the predicted probabilities across the entire test set are less than 0.25. The density of predicted probabilities for each class are shown in Figure 4.9.

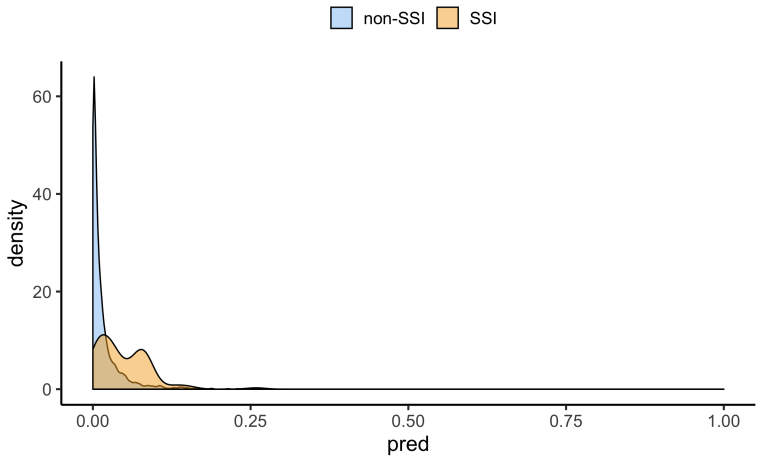

Figure 4.9: A density plot comparing the distribution of the SSI score based on a single model fit to the unbalanced dataset for the SSI and non-SSI training patients.

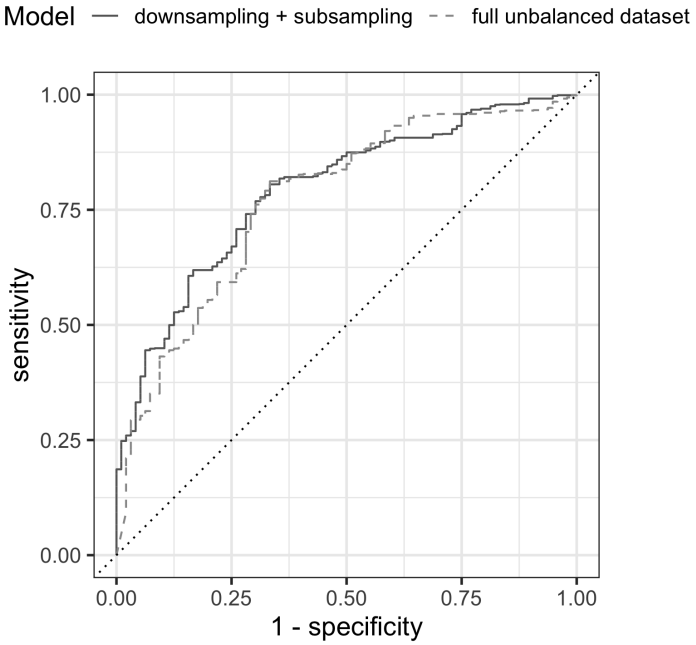

Figure 4.10: A density plot comparing the distribution of the SSI score based on a single model fit to the unbalanced dataset for the SSI and non-SSI training patients.

The correlation between the predictions from the single model based on the full unbalanced dataset and the SSI score (averaged probability predictions across 1000 models) is 0.65. The ROC curve in Figure 4.10 compares these two models.

It might be argued that this single model based on the unbalanced data actually provides a more accurate prediction. After all, is is really plausible that a patient is 70% likely to get an infection? While we don't treat the SSI score as a predicted probability in this thesis, it might still be unintentionally subject to such an interpretation by the medical practitioners who will use it. Perhaps this single model presents a more realistic picture where no individual patient's risk is above $\sim 15\%$.

Figure 4.11 shows the proportion of SSI in the test set patients against their predicted SSI probability from the single model fit to the unbalanced data. This line plot does not have the nice monotonically increasing prediction trajectory that we saw in Figure 4.8 with the aggregated subsampled balanced models. A higher predicted probability does not necessarily reflect an increased SSI rate in the test data when we use the single unbalanced model. This is a very undesirable trait, and we subsequently decide that the aggregated repeated balanced subsampled models provide a more useful SSI score than the single model presented here.

We also tried generating the balanced subsamples within the individual trees in a single forest, rather than generating the downsampled balanced samples for many different forests. The AUC for this approach, however, dropped to 0.7.

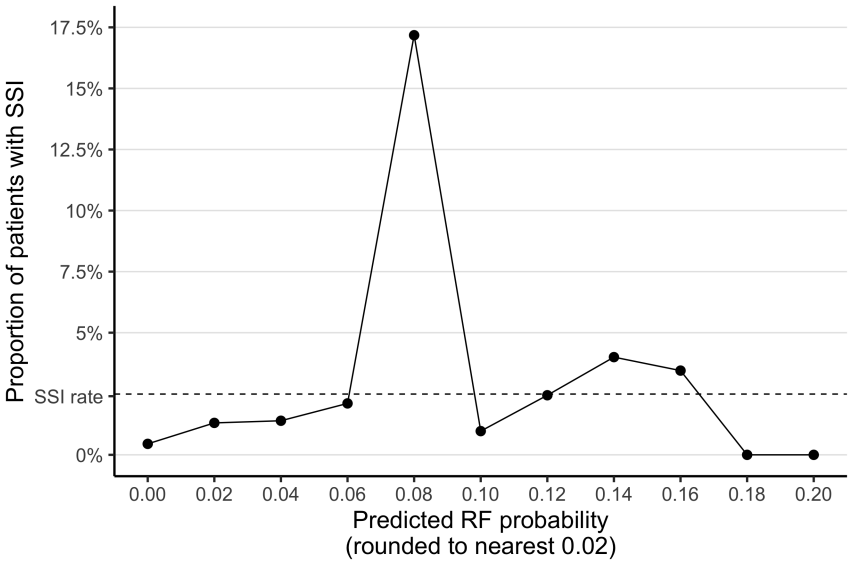

Figure 4.11: A line plot displaying the SSI rate among the test set patients that attained intervals of 0.02 predicted SSI probability from the single unbalanced RF model. The horizontal dotted line corresponds to the overall proportion of SSI in the training data.

Single model fit to downsampled or upsampled dataset

Similarly, we find that if we do a single upsampling or downsampling balancing procedure (i.e. without repeated sampling), the model performance lags behind that of the repeated balanced sampling model. Figure 4.12 displays the density plots for the distribution of predicted SSI probabilities based on the *upsampled balanced* training dataset for the test set patients. Figure 4.13 displays the same plot but for the model based on the *downsampled balanced* training dataset.

Figure 4.14 displays the ROC curves for the up- and down-sampled models (dashed curves) with the ROC curve for the SSI score balanced subsample model. The AUC for the downsampled model is 0.788 and the AUC for the upsampled model is 0.737.

However, these results are just a randomly selected example of downsampling that achieved a similar AUC to the subsampled balanced modeling approach. Figure 4.15 shows the distribution of test-set AUC values across 100 different downsampled models. Most of the downsampled models have AUCs lower than that of the aggregated subsampled balanced model (the orange vertical line).

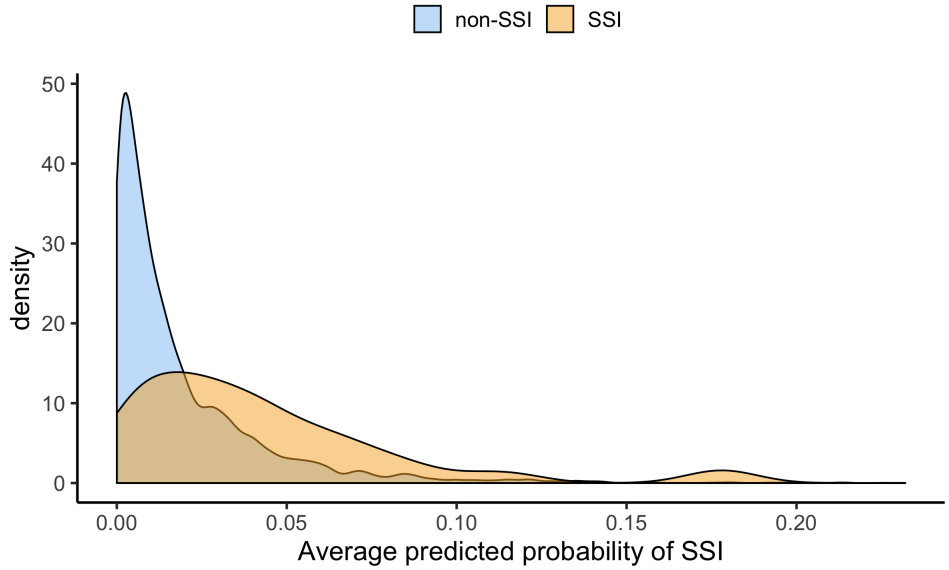

Figure 4.12: A density plot comparing the distribution of the SSI score based on a model fit to a single upsampled balanced dataset for the SSI and non-SSI test patients.

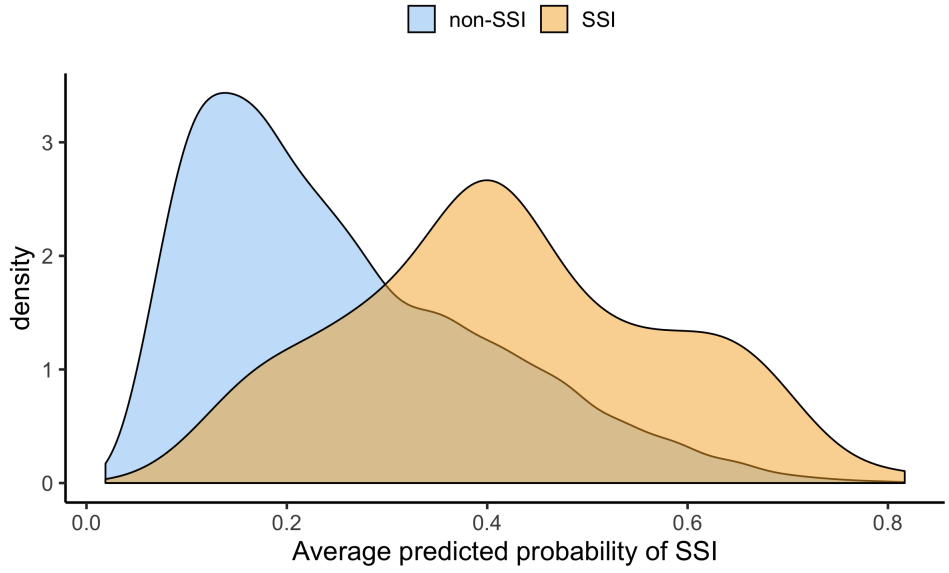

Figure 4.13: A density plot comparing the distribution of the SSI score based on a model fit to a single downsampled balanced dataset for the SSI and non-SSI test patients.

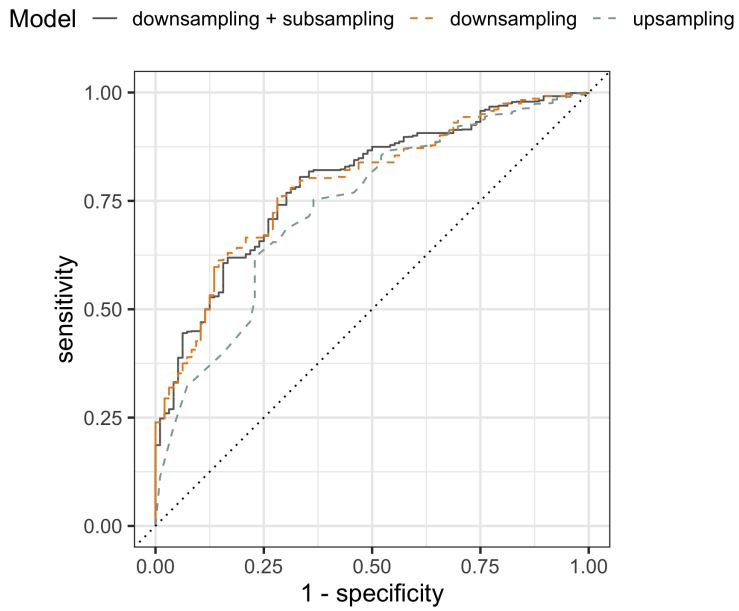

Figure 4.14: A density plot comparing the distribution of the SSI score based on a single model fit to the unbalanced dataset for the SSI and non-SSI training patients.

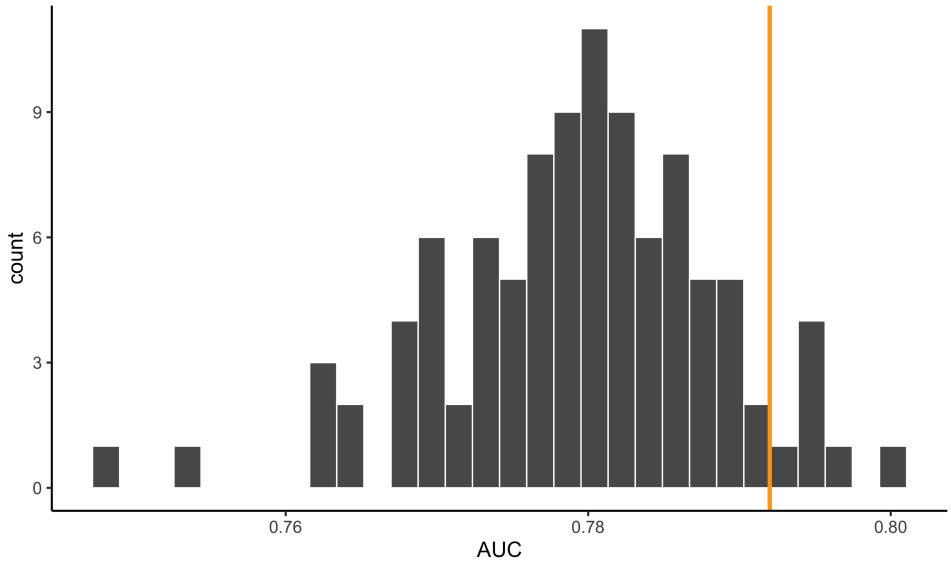

Figure 4.15: A histogram displaying the distribution of test-set AUC values for 100 different downsampled models. The orange line corresponds to the AUC for the aggregated repeated balanced subsample model corresponding to the SSI score.

Modeling procedure categories separately

In Figure 4.16, we separate the ROC curve (for the model based on 15 features) into the risk categories defined by our surgeon collaborators at UC Davis. The model performs the best on patients undergoing procedures with high risk and it performs the worst on patients undergoing procedures with low risk. This is unsurprising, however, since because the patients undergoing high risk procedures naturally have more SSI cases, and the patients undergoing low risk procedures have relatively few SSI cases. Thus, most of the SSI patients in the global unstratified dataset tend to be patients undergoing high risk procedures, and thus the model does a better job at predicting SSI for these patients.

One question that arises from this finding is whether a separate risk-specific models that are fit separately to the patients from each risk group might perform better than a global model fit based on all patients. To answer this question, we compared models fit separately to each risk group with the performance of the global model on each of the risk groups. Figure 4.17 shows that the model fit to just the patients undergoing high-risk procedures performs very similarly to the global model fit to all training patients applied to just the test set patients undergoing high-risk procedures. The global model performs slightly better than the moderate-risk-specific model for the patients undergoing moderate risk procedures, and the global model performs substantially better than the low-risk-specific model for the test set patients undergoing low risk procedures.

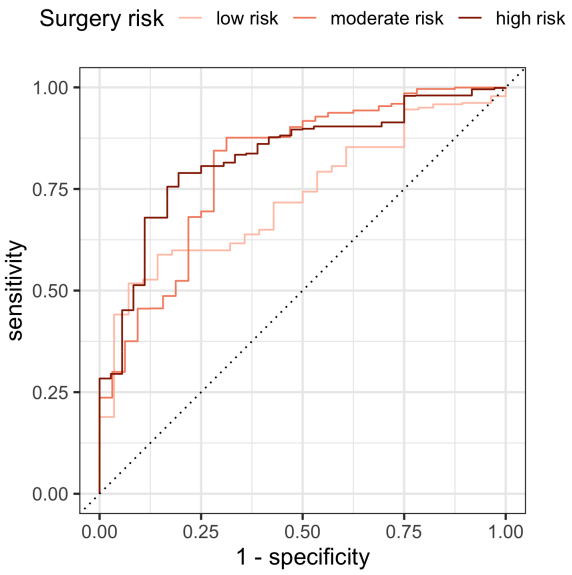

Figure 4.16: The test-set ROC curve for the model built on 15 features separated across procedure risk level. The model performs best on patients with high risk, and worst on patients with low risk.

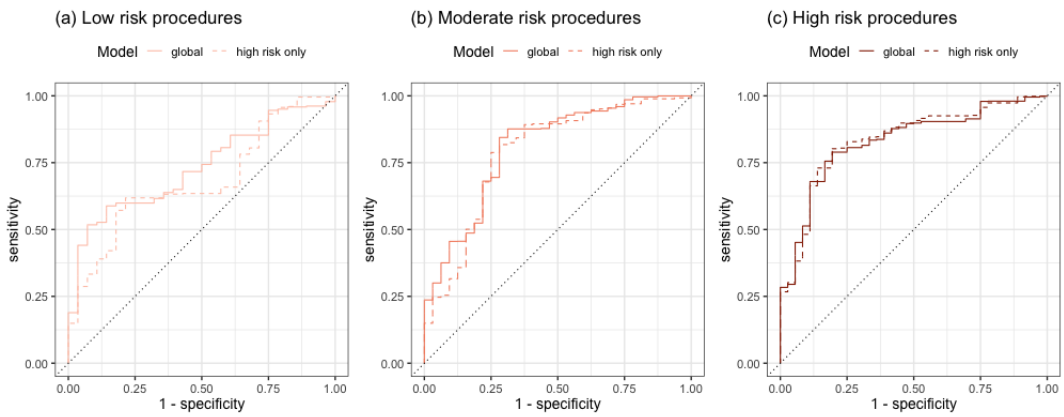

Figure 4.17: The test-set ROC curves for the (a) low, (b) moderate and (c) high risk-specific models and the global models filtered to the patients undergoing procedures of the respective risk-level.

Using NHSN variables only

Finally, recall from Section 4.1 that the existing methods for predicting SSI only use the NHSN patient and surgery features. If we re-fit the subsampled balanced models with only the NHSN variables (i.e. excluding the lab, vitals, and medication data), the test-set predictive performance decreases substantially. Figure 4.18 displays the ROC curves for the NHSN-only model and the model that uses both the NHSN and EHR data (our primary model above). The AUC for the NHSN-only model is 0.71, as compared to the AUC of 0.79 that our NSHN + EHR model obtained.

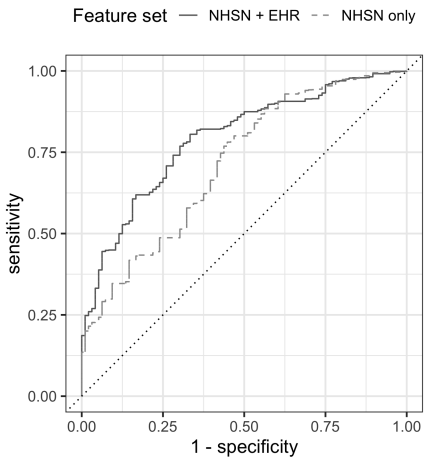

Figure 4.18: The test-set ROC curve for the model built using the top 15 NHSN and lab features and the test-set ROC curve for the model built using just the top 15 NHSN features.

Logistic regression and SVM

When using a logistic regression or support vector machine (SVM) model instead of a random forest model, we find almost identical test set performance. Figure 4.19 displays the ROC curves for the logistic regression version of the model and compares it with the original RF model. The AUC of the logistic regression and SVM models are each also 0.79.

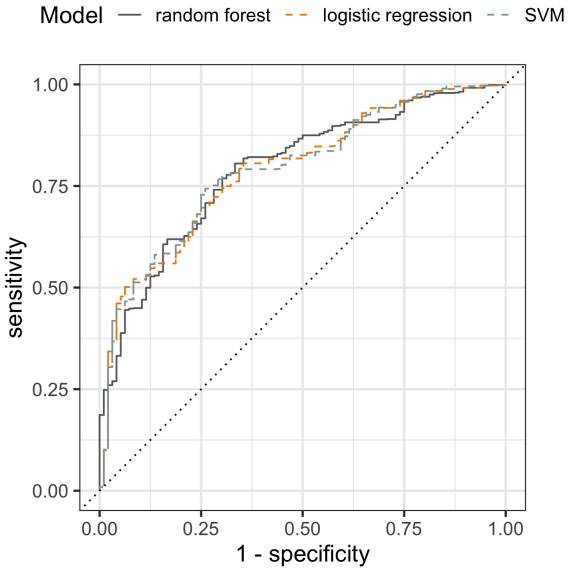

Figure 4.19: The test-set ROC curves for the RF-based SSI model and a logistic regression-based model.

4.8 Conclusion

In this chapter we introduced and implemented a rare-event RF-based model that aggregates the predictions from many subsampled balanced datasets to produce what we call the SSI score. This model predicts Surgical Site Infections 7-days post surgery using both EHR data as well as the NHSN data used by existing methods, and achieves a test-set AUC of 0.79. This SSI model will eventually be validated on an external population and will be implemented at UC Davis for use by clinicians to guide decision-making for SSI prevention. Since the model was built using the population of patients at UC Davis, and has not yet been tested on other populations, its generalizability outside of UC Davis is unknown. However, our modeling approach experiences improved predictive performance when compared with a version that just use the NHSN variables as the existing SSI prediction methods do.

Part III

Causal Inference: Estimating the Effect of Liver Transplant Wait Time on Survival

Chapter 5

The US Liver Transplant Waitlist System

5.1 Introduction

There are currently over 100,000 people in the US waiting for an organ transplant, more than 13,000 of whom are waiting for a liver. Unfortunately, the demand for liver transplants from waitlisted candidates is far greater than the supply of livers from deceased donors. During 2017, only around 8,000 of the more than 13,000 people waiting for a liver received a transplant [27]. While the number of patients undergoing liver transplants is increasing, so too is the waitlist (Figure 5.1). At current rates, the number of people transplanted will never catch up to the number of people waiting to be transplanted. Thus, every time a liver becomes available, a decision needs to be made concerning who gets it.

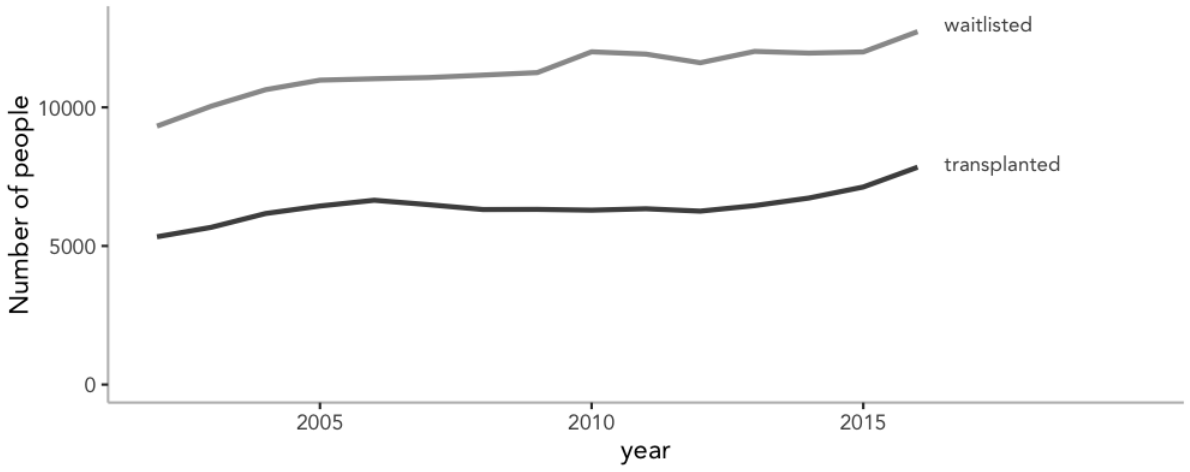

Figure 5.1: The number of people waitlisted and transplanted per year.

Since 2002, livers in the US have been allocated via a "sickest-first" model, where patients who have the lowest estimated three-month transplant-free survival probability are prioritized. However, an unintentional side effect of this the sickest-first system is that patients must be sufficiently ill in order to be eligible to receive a transplant. However, patients who are sicker at the time of transplant may also have diminished post-transplant health outcomes as compared with had they received a transplant at an earlier, healthier stage [88].

There is a clear tradeoff: a patient who is not very sick may live a long time without a transplant (so a transplant is not really needed), whereas a patient who is very sick may not live very long even with the transplant. Ideally, transplants will be assigned to patients at a time such that they will "benefit" the most: i.e. when the length of time they would survive if they receive a transplant "now" is sufficiently greater than the length of time they would survive if they had to wait longer for a transplant.

This concept is known as "transplant-benefit" [50, 68, 81, 88, 56]. A recent survey with 502 participants in Germany indicated that while liver transplant *patients* favored the sickest-first allocation (such as that currently being implemented in the US), all other groups (medical staff, medical students and non-medical university staff and students) favored benefit-based allocation [26]. In contrast, a US-based study found that there was little support amongst members of the public that allocation decisions should be based solely on the sickest-first criteria, and they found a general support for maximizing outcomes after transplantation [75].

In this thesis, our goal is not to estimate transplant benefit for individual patients, but to examine whether such a survival benefit exists at all, and if so, how much of a benefit can be expected from receiving a transplant, say, one month earlier. Our unique analytic approach makes use of the fact that due to donor-recipient blood type matching, individuals with certain blood types have shorter transplant wait-times than similar individuals with other blood types, and we position blood type as a causal instrument [7, 6]. Since similar individuals with different blood types (the instrument) have different wait times (the treatment), but blood type should have no effect on survival (except via its effect on wait time), any survival differences we observe between patients with different blood types must be due to the difference in wait times. The foundation of this argument is based on the idea that there is nothing *else* that blood type is related to (other than wait time, that is) that also impacts post-transplant survival.

Our findings indicate that if everyone received a transplant 6 months earlier then there would be a 4.2% reduction in death by only 24 months. Thus, we conclude that there is indeed a survival benefit that arises from receiving a transplant earlier. This implies that a benefit-based allocation system might be very effective at increasing overall post-transplant survival, however, to actually implement this would involve developing a method that can reliably estimate the benefit for individual patients, a feat that is outside the scope of this thesis.

This Chapter will introduce the current liver transplantation system in the US, explore the individual-level dataset provided by UNOS for all waitlisted patients in the US, and will identify some criticisms of - and alternatives to - the current allocation system. Chapter

6 will discuss and implement an instrumental variables analysis (with blood type as an instrument) to estimate the average effect of increasing/decreasing wait time on survival.

5.2 Liver transplantation in the USA

UNOS: a country-wide organ allocation organization

In the United States, the allocation of livers (and other organs) for transplantation is administered by a private non-profit organization called the United Network for Organ Sharing (UNOS). In 2002, UNOS adopted a Model for End Stage Liver Disease (MELD) score-based system for allocating livers to candidates across 11 distinct geographical regions [46]. Each region is managed by a region-specific Organ Procurement Organization (OPO). Under this system, each candidate is placed on a waitlist that is specific to (1) which of the 11 geographic regions they reside, and (2) their blood type (candidates can only accept livers from donors with compatible blood types). When a donor liver becomes available, it is offered first to the candidate with the highest MELD score who has a compatible blood type and is registered in the same geographic region from where the donor came [67].

The MELD score

The Model for End-Stage Liver Disease (MELD) score was developed during a study based on a set of 231 patients across 4 US medical centers who had undergone a Transjugular Intrahepatic Portosystemic Shunt (TIPS) procedure [65]. The original authors found the MELD score to be correlated with three-month survival among this specific set of patients. Despite the relatively small sample size and the specific set of patients on which it was developed, over the next few years the MELD score was shown by other researchers to be a reasonable predictor of three-month transplant-free survival for patients with a wide range of liver problems (not just those who had undergone a TIPS procedure) [67, 107, 46].

The MELD score can take values that range from 6 (less ill) to 40 (extremely ill), with values calculated to be below 6 being rounded up to 6, and values calculated to be above 40 being rounded down to 40. The score is based on measurements of serum bilirubin (mg/dL), serum creatinine (mg/dL), and the international normalized ratio for prothrombin time (INR) and is calculated using the following formula:

$$\text{MELD} = 6.43 + 3.78\ln(\text{serum bilirubin}) + 11.2\ln(\text{INR}) + 9.57\ln(\text{serum creatinine})$$

So that the resulting score is an integer, the output of the MELD formula is rounded to the nearest whole number.

The three-month observed mortality by MELD score is shown in Table 5.1 [108]. By allocating each liver to the patient with the highest MELD score, the current system prioritizes allocating donor livers to the sickest patients first (those who have the lowest estimated chance of three-month transplant-free survival).

MELD score	Observed mortality
40 or more	71.3%
30-39	52.6%
20-29	19.6%
10-9	6.0%
<9	1.9%

Table 5.1: The observed three-month transplant-free mortality by MELD score from [108].

Geographical matching

The US is split into 11 regions (Figure 5.2). Each region is further split into 58 sub-regions managed by separate Organ Procurement Organizations (OPOs) (Figure 5.3). These OPOs approximately cover individual states, although some states have multiple OPOs (such as California which has OPOs 55, 56, 57, and 58), and other OPOs cover multiple states (such as OPO 52 which covers much of Colorado and Wyoming). When an organ becomes available, the allocation algorithm does not initially consider all waitlisted patients in the entire country, but rather begins by looking first for recipients locally in the same OPO, and then if no suitable recipient is found in the OPO, the search is expanded to the entire region from which the organ came, and then eventually to the entire country.

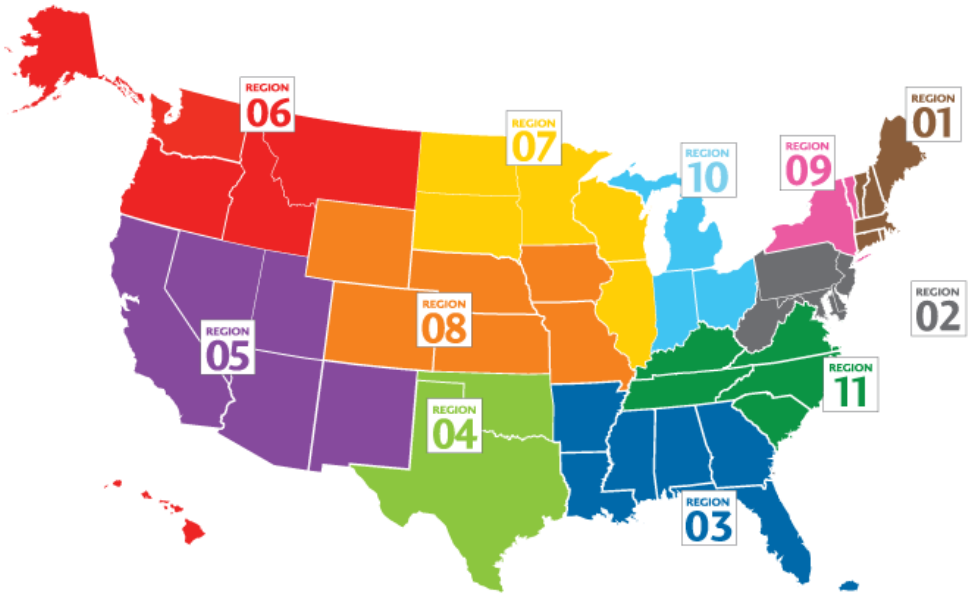

Figure 5.2: A map of the 11 UNOS regions sourced from the UNOS website.

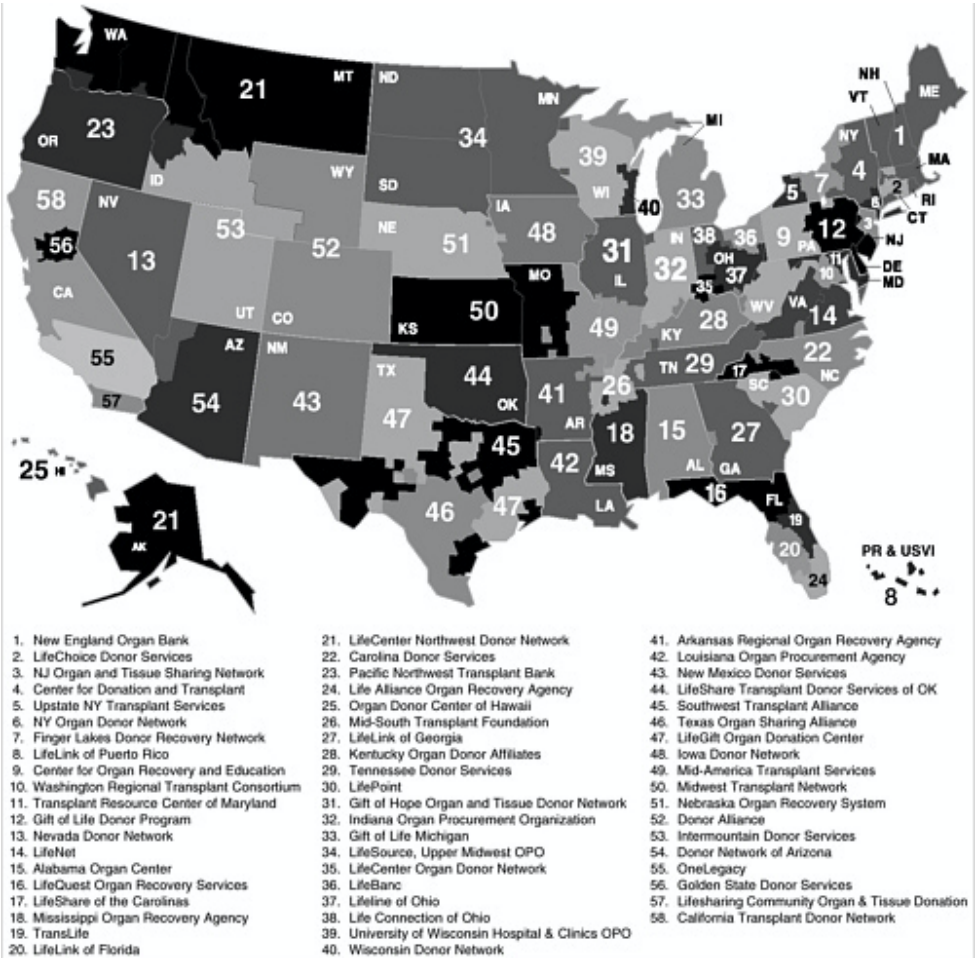

Figure 5.3: A map of the 58 OPOs sourced from [76].

Blood-type matching

Donor livers can only be transplanted into recipients with a compatible blood type. The donor-recipient blood type compatabilities and frequencies are described in Table 5.2.

Recipient blood type	Frequency	Acceptable donor blood type
O	46%	O only
A	37%	A and O
B	12%	B and O
AB	4%	A, B, AB and O

Table 5.2: The frequency of each blood type among the recipients, and the donor blood types that they are compatible with.

Recipients with blood type AB are universal recipients: they can receive livers from any donor blood type. At the other end of the spectrum, *donors* with blood type O are universal donors: they can provide livers to any recipient blood type. As we will discuss in Section 6.3, this means that recipients with blood type AB end up with shorter wait times (because they can receive livers from donors with any blood type), and recipients with blood type O end up with longer wait times (because type O livers, the only type that they can receive, are often sent to recipients with other blood types).

MELD exceptions

There are several disease-based exceptions to the usual MELD score system. For example, patients with hepatocellular carcinoma receive extra MELD points, placing them higher on the waiting list than their calculated MELD score implies.

The following conditions are automatically assigned a MELD Score of 22 (28 in case of hyperoxaluria), with a 10% increase in score every 3 months from diagnosis.

- Hepatocellular carcinoma (HCC) with one lesion between 2-5cm or two to three lesions ≤ 3 cm (Milan criteria), provided no vascular invasion or extrahepatic disease.
- Hepatopulmonary syndrome with PaO2 ≤ 60 mmHg on room air.
- Portopulmonary hypertension, with mean pulmonary artery pressure (mPAP) ≥ 25 mmHg at rest but maintained ≤ 35 mmHg with treatment.
- Hepatic artery thrombosis 714 days post-liver transplantation.
- Familial amyloid polyneuropathy, as diagnosed by identification of the transthyretin (TTR) gene mutation by DNA analysis or mass spectrometry in a biopsy sample and confirmation of amyloid deposition in an involved organ.
- Primary hyperoxaluria with evidence of alanine glyoxylate aminotransferase deficiency (these patients requires combined liver-kidney transplantation).
- Cystic fibrosis with FEV1 (forced expiratory volume in 1 second) $\leq 40\%$.
- Hilar cholangiocarcinoma.

Moreover, patients with acute (sudden and severe onset) liver failure and a life expectancy of hours to a few days without a transplant are placed in a special category known as Stats 1A (or 1B if the patient is under 18 years of age). These Status 1 exceptions are prioritized by the MELD system.

Since the normal MELD-based allocation trajectory does not apply to patients with exceptions, in this thesis, we ignore all patients who received MELD exceptions.

Breaking ties

While wait time is not an explicit component of the allocation algorithm, it does play a role as a tie breaker. If there are two or more patients with the same blood type and the same MELD score in the geographic region under consideration, then the liver will be offered to the patient who has had the longest wait time.

Calculating the wait time, however, is not quite as simple as time since listing. Instead, wait time is calculated within tiers. The MELD score is broken down into levels:

- greater than or equal to 25
- 19-24
- 11-18
- less than or equal to 10

As the patient's MELD score moves them up to a new level, a new wait time clock starts. However, if a patient moves backwards to a lower MELD score level, the waiting time accumulated at the higher score remains (but not the other way around).

5.3 The UNOS STAR waitlist dataset

The United Network for Organ Sharing (UNOS) hosts and makes available a wide range of the data collected across the national transplant system. While most of the data that UNOS makes available to the public is aggregated (e.g. by recipient state or recipient age), UNOS also provides patient-level data by request in the form of the Standard Transplant Analysis and Research (STAR) dataset. The STAR data that we obtained from UNOS contains patient level information for over 260,000 liver waiting list candidates who were waitlisted starting in 1986 through to the end of 2016. Of these candidates, almost 150,000 had received a transplant by the end of 2016. In this thesis, we do not consider all of these patients. We exclude the following patients:

- Patients waitlisted before February 27, 2002: the date that the MELD-based allocation system began.
- Patients who received a MELD exception or Status 1.
- Patients who are 18 years or younger: children under the age of 18 have a different waitlist based on a PELD (Pediatric End-Stage Liver Disease) score.
- Patients undergoing multiple transplants.
- Patients receiving partial liver transplants (as opposed to whole organ transplants).

- Patients receiving transplants from living donors (as opposed to deceased donors), since transplants through living donors are not based on the wait list system.
- Patients who have had previous liver transplants.

After applying these filters, we have 64,251 patients remaining in our cohort. The original dataset has 394 columns, but we restrict just 43 variables deemed relevant to our study. These variables are described in Table 5.3.

In this section, we will explore the STAR dataset, and the examine how wait time, the MELD score, and survival are all related to one another.

Variable	Description
WL_ID_CODE	Waitlist ID code
INIT_DATE	Date of listing
GENDER	Gender
ABO	Blood type
INIT_AGE	Age
ETHCAT	Ethnicity
WORK_INCOME_TCR	Work income
EDUCATION	Education level
WGT_KG_TCR	Weight (kg)
HGT_CM_TCR	Height (cm)
BMI_TCR	BMI
PRI_PAYMENT_TCR	Insurance type
INIT_MELD_PELD_LAB_SCORE	Initial meld score
DGN_CODE	Diagnosis category
DIAB	Diabetes status
FUNC_STAT_TCR	Health status
PREV_AB_SURG_TCR	Prior abdominal surgery
BACT_PERIT_TCR	Bacterial Peritonitis
MALIG_TCR	Malignancies
TIPSS_TCR	TIPSS procedure
HBV_CORE	HBV infection
HCV_SEROSTATUS	Hepatitis C infection
EBV_SEROSTATUS	EBV infection
HIV_SEROSTATUS	HIV infection
CMV_STATUS	CMV infection
INIT_OPO_CTR_CODE	OPO
REGION	Geographic region
PERM_STATE	State
TX_DATE	Date of transplant
ABO_DON	Blood type of donor
COMPOSITE_DEATH_DATE	Death date if after transplant
DEATH_DATE	Death date if during waitlisting
END_DATE	Date of final recorded information
FINAL_MELD_PELD_LAB_SCORE	Final recorded MELD score
GRF_FAIL_DATE	Date of graft failure
REM_CD	Removed from waitlist
SHARE_TY	Local, regional, or international donor
COLD_ISCH	Cold ischemic time
DEATH_MECH_DON	Cause of donor death
AGE_DON	Age of donor
BMI_DON_CALC	BMI of donor

Table 5.3: The data dictionary for the 43 variables from the STAR dataset.

Unsurprisingly most liver transplant patients are older, with most transplant recipients being in their 50s (Figure 5.4).

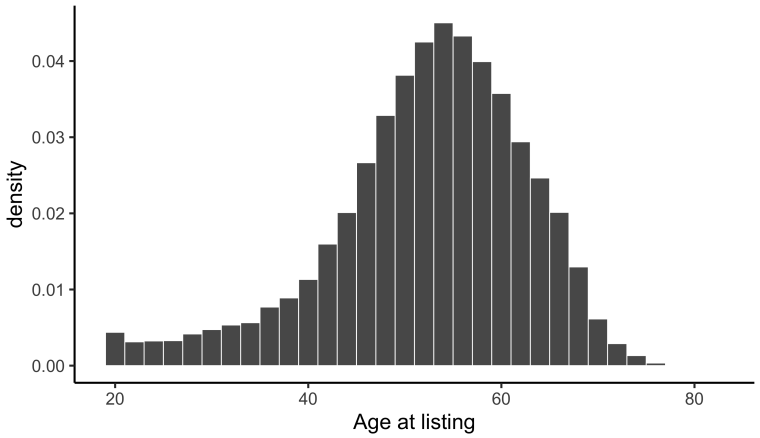

Figure 5.4: A histogram displaying the distribution of age at the time of listing.

There is a wide range of MELD scores at which patients are initially listed on the waitlist (Figure 5.5). While most patients are listed at a MELD score that is under 20, many people are listed at MELD scores over 20 where the 3-month mortality is estimated to be over 20% (Table 5.1). Moreover, Figure 5.6 shows that there are some patients even with these higher listing MELD scores are waiting years for a transplant, though, as might be expected, wait time does tend to decrease with MELD score at listing.

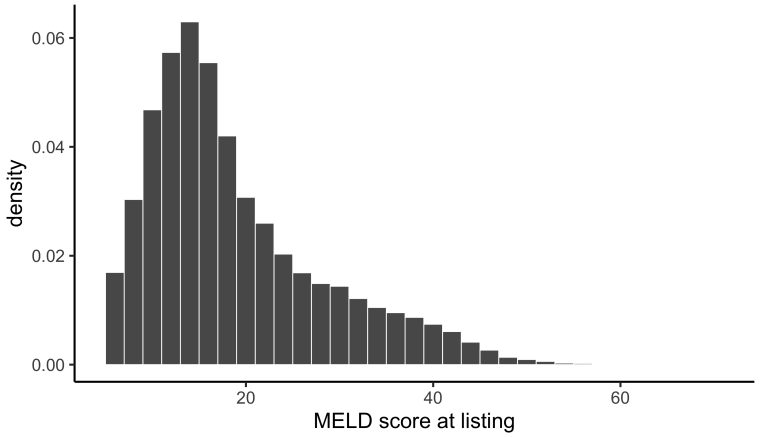

Figure 5.5: A histogram displaying the distribution of MELD score at the time of listing.

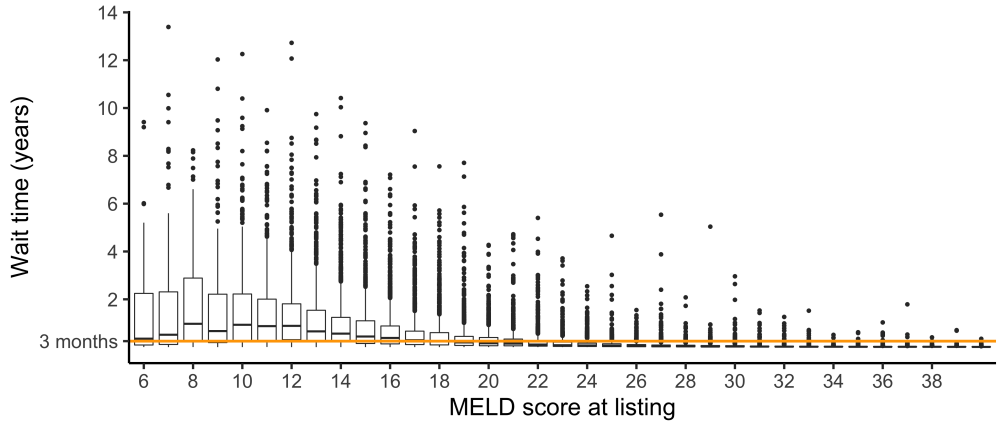

Figure 5.6: Boxplots displaying the distribution of wait time (in years) as a function of initial MELD score at listing. The orange line represents three-months.

Figure 5.7 displays the proportion of patients that died on the waitlist based on MELD score at listing. The MELD category with the highest risk of death on the waitlist is between a MELD of 8 and 15 or so, probably due to the fact that these patients have to wait so long for a transplant. The risk of waitlist death seems to increase until around a MELD score of 12, after which the risk decreases until a MELD score of around 35, at which point the risk of waitlist death starts to increase again (which makes sense, since these patients are inherently very sick).

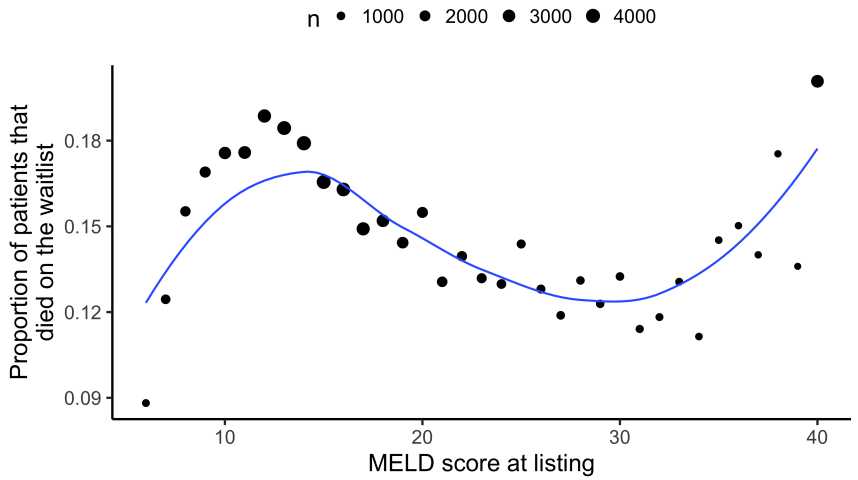

Figure 5.7: A scatterplot displaying the proportion of patients who died on the waitlist by MELD score at listing. The size of the point corresponds to the number of patients with each MELD score.

Correspondingly, Figure 5.8 shows that as the MELD score at transplantation increases, the risk of death within three-months post-transplantation also increases. This means that patients who get transplanted at higher MELD scores are also dying at higher rates post-transplantation. This is one of the key arguments against the sickest-first allocation algorithm. This is a particularly concerning finding since the average transplant MELD score has been increasing over time as shown in Figure 5.9.

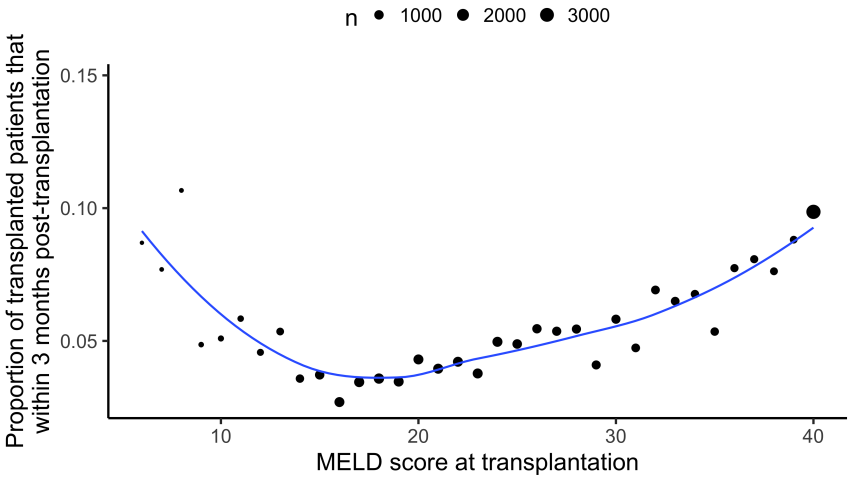

Figure 5.8: A scatterplot displaying the proportion of patients who died within three months-post transplantation by MELD score at listing.

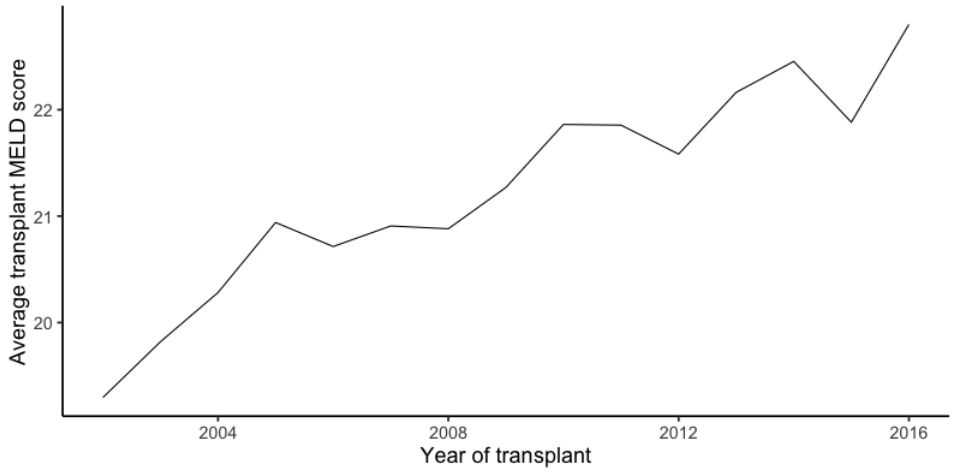

Figure 5.9: A line plot displaying the increase in average MELD score at transplantation over time.

5.4 Criticisms of MELD

While the MELD score continues to be the primary method for liver allocation, there have been many studies drawing criticisms on the effectiveness of the MELD score. For instance, several studies have shown that MELD is a poor predictor of post-transplant survival [54, 45], that the MELD score should include additional relevant lab values such as serum sodium [86], that the MELD score does not induce equal transplant opportunities across race and sex [70], and that patient prognosis for some diseases are better captured by MELD than others [10].

One of the most influential criticisms is that wait times are dramatically different across different parts of the country [113]. Even within regions, different OPOs have vastly different wait-times. Figure 5.10 displays the proportion of patients who have been transplanted within three months of listing in each state. It is clear that Southern states have shorter transplant wait-times (almost half of the patients are transplanted within 3 months of listing) than states in the North East, along the West Coast, and in the Midwest (which typically have fewer than a third of their patients transplanted within 3 months).

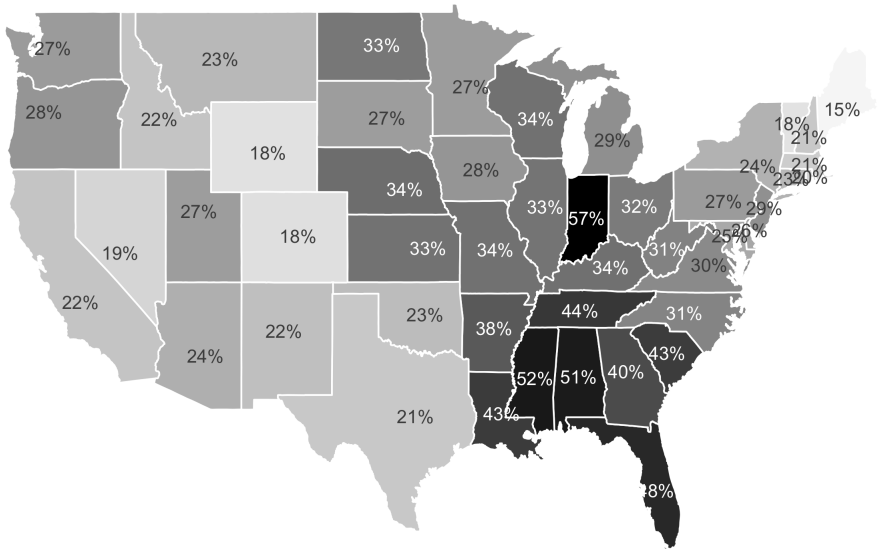

Figure 5.10: A map that displays the proportion of patients listed between Jan 1 2015 and Dec 31 2015 who were transplanted within 3 months of listing in each state.

This difference in wait time could be due to a number of different factors [98]; including

- Patients are listed at different initial MELD scores in different states
- Fewer livers being donated in some states versus others

- More people on the waiting list in some states versus others
- People in some states are sicker when they are placed on the waiting list
- Doctors in some states giving more MELD exceptions.

Figure 5.11 shows the average wait time in days to transplant against the average initial MELD score for each state.

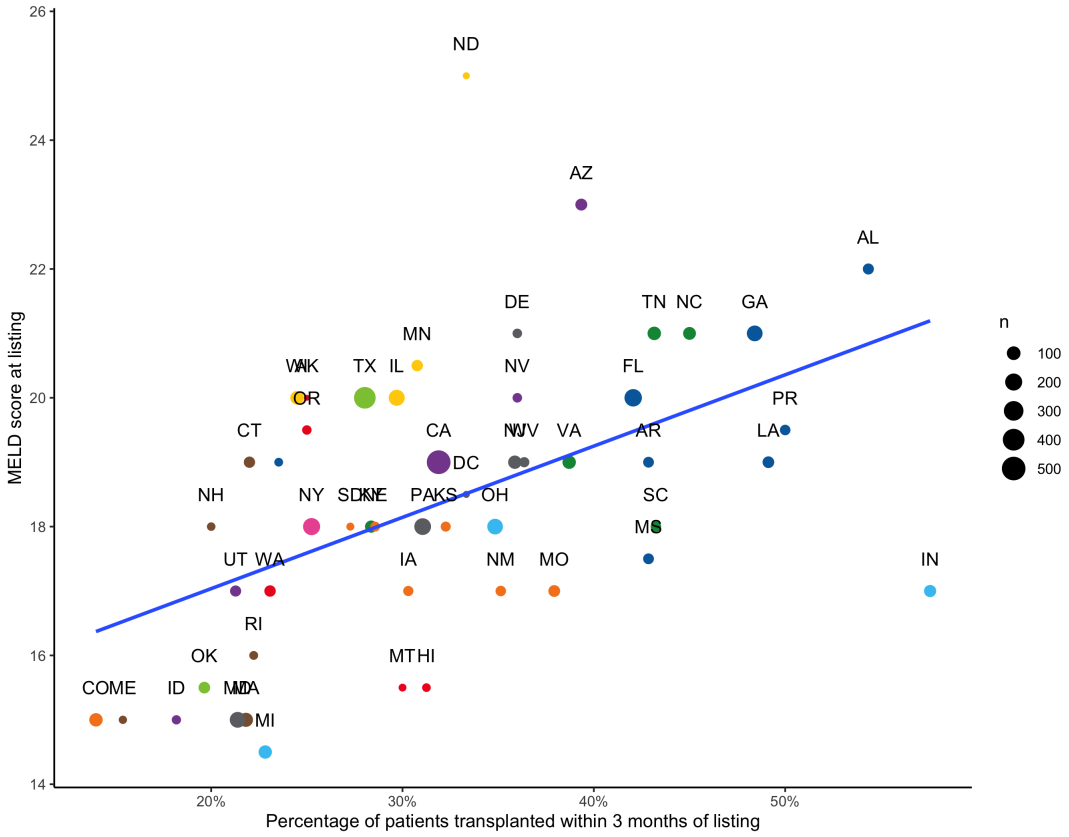

Figure 5.11: A scatterplot displaying the proportion of patients listed between Jan 1 2015 and Dec 31 2015 who were transplanted within 3 months of listing in each state against the state's average initial MELD score at listing.

5.5 Alternatives to MELD-based allocation: survival benefit

So long as there continues to be a shortage of donor organs available (we cannot yet synthetically create organs), it is important to ensure that the organs that are available are distributed reasonably. There have been many studies examining the potential effectiveness

of alternative allocation schemes [73]. The most common schemes discussed are based on one or more of the following principles:

- Equity: equal chance of transplantation such as first-come first-served
- Utility: organs are allocated to the recipient who is likely to have the best outcome
- Benefit: organs are allocated to the patient who has the greatest benefit, so taking into account the risks of dying with and without a transplant;
- Urgency: to reduce the risk of dying on the list;

The current sickest-first MELD-based allocation system is based on urgency. Much of the recent literature, however, has focused on allocation systems based on benefit [68, 56, 81, 88, 101]. Under the most common formulation of the benefit-based system, an organ should be allocated to the patient estimated to experience the biggest difference in survival with and without the organ.

The most common formulation of a benefit-based allocation procedure is based on the benefit associated with the two competing scenarios: "the patient receives a transplant now" versus "the patient doesn't receive a transplant at all". However, this is not the scenario that most patients face. If a patient does not receive the currently available organ, they will most likely receive a different organ in the near future, rather than not receive any organ at all.

In addition, each of these studies each estimate benefit by comparing patients who received a transplant to patients who did not. Additional issues arise when the individuals being considered who did not receive an organ are also the individuals who died on the waitlist. These patients are fundamentally different to those who did receive an organ. The problem is that these patients died before they had a chance to receive a transplant, rather than the transplant was withheld from them.

Instead of defining survival benefit as the difference between survival with a transplant and transplant-free survival, we posit that it would be more meaningful to estimate the extent to which a patient will benefit if they were to receive a transplant *now* versus *later*. This quantity is much more relevant to the decision being made when a transplant becomes available: should the patient receive *this* liver, or some liver that will become available at a *later* date (rather than: should the patient receive this liver or no liver at all). This is the definition of survival benefit that we will focus on in this thesis.

Another issue that each of the existing works face is defining survival time. All authors define survival time as time from initial listing to death. However, both their likelihood of transplantation and their survival are impacted by how sick they were at time 0, as measured by the MELD score. Sickness level is thus inbuilt into their measure of time. Ideally, all patients will be on an even playing field at time 0: this means that they each have similar levels of sickness, and they will all have equal probabilities of treatment (transplantation). Thus, in this thesis we will treat time 0 as the first time the patient attains a specific MELD score (e.g. MELD 18). We will discuss this idea further in Section 6.2.

5.6 Conclusion

This chapter introduced the current status of the liver transplant allocation procedure and the impacts of this procedure on the patients who are awaiting transplantation on the waitlist.

The current allocation procedure is based on urgency and gives the available liver to the "sickest patient first", as determined by the MELD score. We examined several issues with the MELD score allocation procedure, including geographic disparities in transplant availability, and the fact that patients need to get sufficiently sick before they can receive a transplant, which is at odds with the fact that transplants that take place for patients with higher MELD scores lead to decreased post-transplant survival.

With these criticisms of the current system in mind, we introduced the idea of allocating organs based on survival benefit, and discussed the shortcomings of current methodology intending to implement or model various versions of survival benefit allocation systems.

In the Chapter 6, we will provide an instrumental variables-based estimate of the average survival benefit experienced by receiving a transplant one (or any specific number) of months earlier. We will show that the survival benefit does exist, but that further research would be needed to develop a method that can reliably estimate this benefit for individual patients on the waitlist.

Chapter 6

Estimating Survival Benefit using Blood Type as an Instrument

6.1 Introduction

In the previous chapter we introduced the liver transplant allocation system as it is currently implemented in the United States by UNOS, along with the STAR individual-level dataset provided by UNOS on all transplant waitlist patients. We identified a range of issues with the current allocation system, and introduced the idea of allocating organs based on survival benefit. In this chapter, we will use a unique instrumental variables approach to estimate average survival benefit using blood type as an instrument.

First, we need to position our problem in the context of causal inference, where the goal is to estimate the effect of a treatment on an outcome. In our case, the treatment that we consider is wait time to transplantation, and the outcome is survival. As an example, if a patient would survive for 5 years if they had to wait 6 months for a transplant, but they would survive for 9 years if they got a transplant today, then their 6-month transplant survival benefit is 4 years (since they would survive for 4 years longer if they received a transplant 6 months earlier).

If we could observe the survival of a patient under all possible transplant wait times, then we could come up with a curve that showed survival time as a function of wait time as imagined by Figure 6.1. Unfortunately, due to the fundamental problem of causal inference, we can only ever observe a single point on this curve, corresponding to the actual wait time experienced and the subsequent post-transplant survival time.

If we could observe the relationship between wait time and post-transplant survival for patients "equivalent" to our patient of interest, then we could fill in the above hypothetical curve with the observations made on similar patients with different wait times.

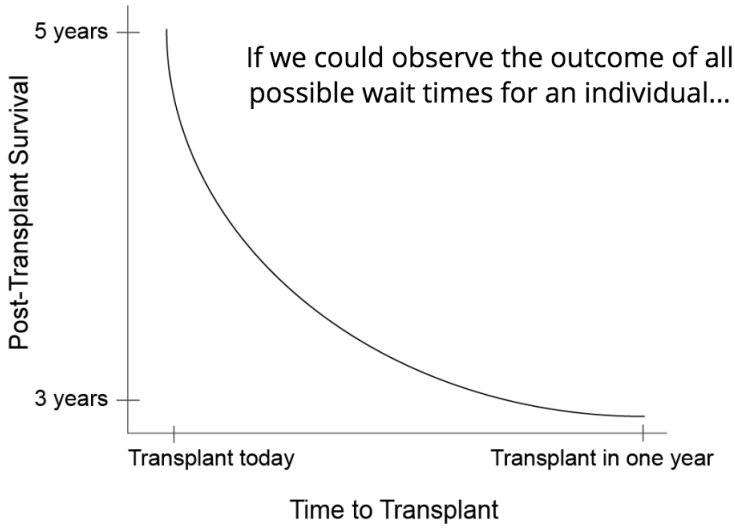

Figure 6.1: A hypothetical causal curve for an individual patient that shows their post-transplant survival as a function of wait time to transplantation. Unfortunately due to the fundamental problem of causal inference, and the restricting laws of reality, we only ever observe a single point on this curve (the actual wait time experienced and the subsequent survival time following transplantation)

However, the differences between those who received a transplant earlier and those who received a transplant later are profound: those who were transplanted earlier are more likely to either be sicker or get sicker faster than those who were transplanted later. Sickness is a confounder: it affects both the treatment and the outcome.

Estimating the effect of the transplant wait time on survival using this observational data is thus a very difficult problem: it is impossible to claim directly that any difference in survival between those transplanted sooner vs later is due to the difference in transplant wait time rather than any of the other differences that exist between these two groups of individuals.

As discussed in Section 5.5, all previous studies looked at estimating the benefit in terms of survival of *receiving a transplant now* versus *never receiving a transplant at all*. However, since all patients on the waitlist will eventually receive a transplant if they live long enough, a more relevant quantity is the benefit of *receiving a transplant now* versus *receiving a transplant later*, as we consider in this thesis.

If the typical wait time for the next donor liver to become available is 20 days, then a benefit-based allocation algorithm would be one for which the patient deemed to have the highest 20-day survival benefit is offered the transplant first.

While estimating the actual transplant benefit for individual patients is an incredibly difficult task, the task of this thesis is to first ask how much of a survival benefit exists when receiving a transplant a month earlier when averaged across the waitlisted population. Even

the task of calculating the *average* transplant benefit is an incredibly difficult one, since there is at least one confounder that we know of that we cannot measure (rate of increasing sickness). A confounder is a variable (measurable or not) that affects both the treatment variable and the outcome variable [36].

In this chapter, we first identify sources of confounding and describe how we alleviate them: by both defining time in a sensible way, and by using a method called Instrumental Variables (IV) [6]. We will spend much of this chapter setting up the IV analysis that forms the basis of this project, and eventually present the results that show that if everyone received a transplant 6 months earlier, then the average reduction in death rate by 24 months would be 4.2%.

6.2 Identifying confounders

There is one obvious *measurable* confounder in this study: the MELD score at listing. A higher MELD score means that transplantation will happen sooner, but also that the patient is sicker. Thus MELD influences both the treatment (time to transplant) and the outcome (survival). In order to remove this confounding factor from play, we decided to measure time from the first time a patient achieves a specific MELD score, such as MELD 18. Every patient thus starts at time 0 on equal footing in the eyes of transplantation and mortality. This is in contrast to the prior work discussed in Section 5.5, each of which began time at the time of listing, and did not adequately deal with the MELD confounding [68, 56, 81, 88, 101].

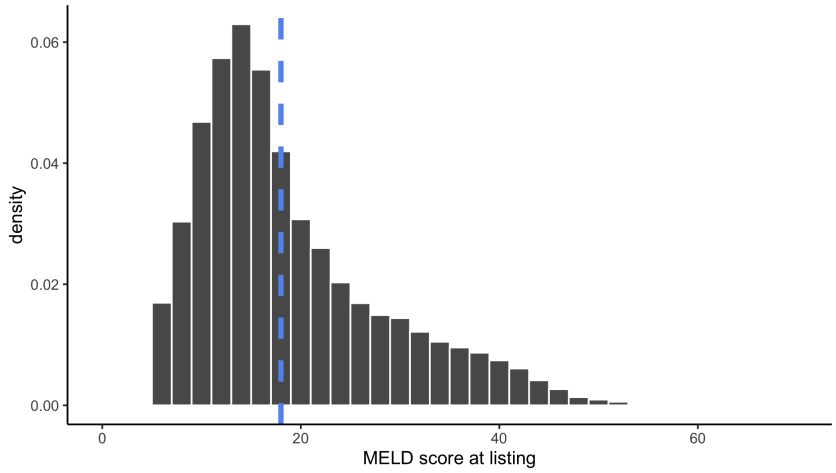

Figure 6.2: A histogram displaying the distribution of initial MELD score across all patients listed since Feb 27 2002 (the date of UNOS' introduction of the MELD score). A vertical line represents a MELD score of 18.

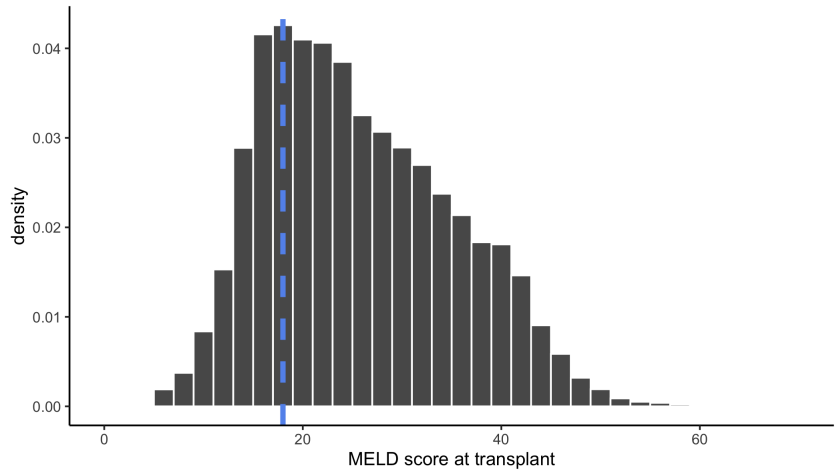

Figure 6.3: A histogram displaying the distribution of MELD score at transplantation across all patients listed since Feb 27 2002 (the date of UNOS' introduction of the MELD score). A vertical line represents a MELD score of 18.

However, there is a drawback of setting time 0 to be MELD 18: we can only capture the patients who at some point had a MELD score of 18. This means that their MELD score at listing must be 18 or less, and their MELD score at transplantation must be 18 or over (Figures 6.2 and 6.3). Of the entire set of eligible waitlisted patients, 13,575 (21.7%) of them satisfy this criteria, and a MELD score of 18 is the value that captures the most patients. Later, we will test our final conclusions to check whether they hold when we consider time 0 to correspond to different MELD scores.

While we have removed the observable confounder of current MELD score by defining time 0 as MELD 18, there is one more confounder that we have unfortunately not removed: the *rate* at which each patient gets sicker after time 0. The patients who get sicker faster will be transplanted earlier, but are also more likely to die earlier. Unfortunately, at time 0, we do not know what rate each patient will get sicker. Thus rate of sickness increase is not only a confounder, but it is an *unmeasured* confounder.

Since our confounder is unmeasured, traditional causal effect estimation methods for dealing with confounders in observational studies, such as conditioning on the confounder, matching [85], and stratification [28] will be unable to remove the bias.

Fortunately, there are sometimes features of the underlying experimental design that allow us to sidestep the unobserved confounder. For instance, if you can find a variable, called an "instrument", that is highly correlated with the treatment but satisfies the *exclusion restriction* (that it is not related to the outcome in any way other than through the treatment), then you can use this instrument as a quasi-treatment variable to get an unbiased estimate of the causal effect. This method is called Instrumental Variables (IV) [6].

If you do believe the exclusion restriction, then if you can estimate the effect of this instrument on the outcome, some portion of that effect will also be capturing the uncon-

founded effect of the treatment on the outcome. The nice thing about IV is that not only will it deal with the unmeasured confounder we *do* know about, it will also deal with any potential unmeasured confounders that we *don't* know about.

The instrument that we will work with in this thesis is *blood type*. Showing that the instrument is correlated with the treatment is easy. The hard part is coming up with a convincing enough argument that the exclusion restriction holds.

6.3 Blood type as an instrument

In our transplant setting, a particularly nice instrument is **blood type**. As we described in Section 5.2:

- A donor with **blood type O** is a universal donor: can donate to O, A, B or AB
- A donor with **blood type A** can donate to A or AB
- A donor with **blood type B** can donate to B or AB
- A donor with **blood type AB** can only donate to AB

This is summarized in Figure 6.4. These rules are confirmed in the data (Figure 6.5), although we do see that livers from donors with blood type A are occasionally given to type-O recipients.

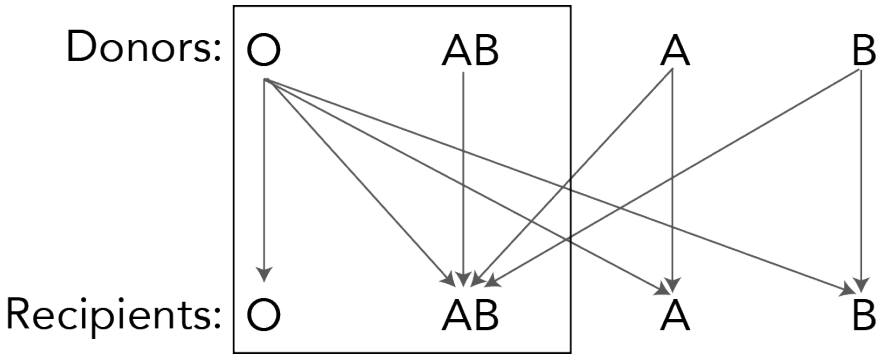

Figure 6.4: A diagram displaying how recipients with blood type AB have a larger pool of potential donors than do recipients with blood type O. An arrow from a donor blood type to a recipient blood type implies that donors with the specified blood type can donate organs to the corresponding recipient blood type.

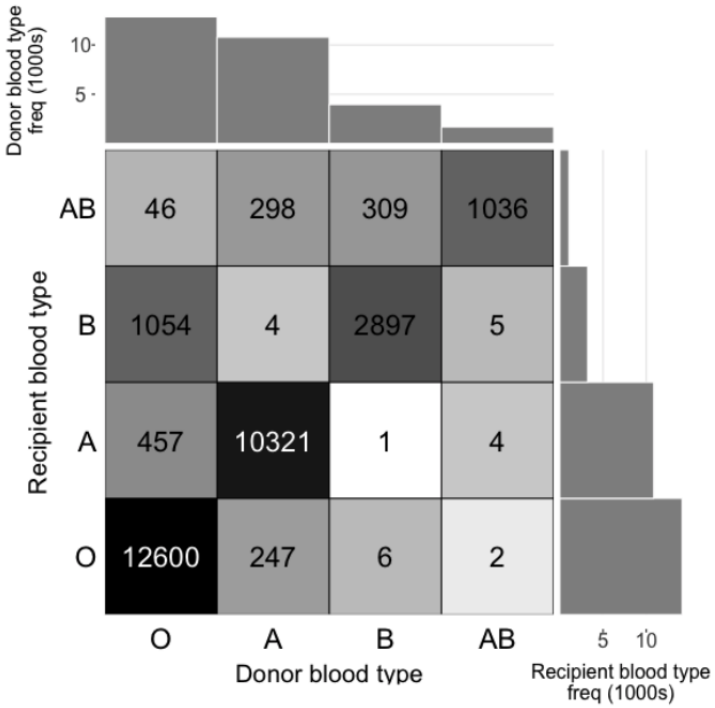

Figure 6.5: A superheatmap displaying the distribution of donor-recipient blood type combinations in the STAR dataset.

Type AB recipients are *universal recipients*, while Type O donors are *universal donors*. Type O recipients can only receive a liver from a Type O donor, but Type O donor livers are often sent to patients other blood types. Simultaneously Type AB recipients can receive livers from donors with any blood type. The result is that patients with blood Type O often need to wait longer for a transplant than patients of a similar sickness level with blood Type AB.

To justify that blood type is indeed a reasonable instrument for this problem, we need to justify a few things:

1. **Relevance**: the instrument (blood type) is correlated with the treatment (wait time)
2. **The exclusion restriction**: the instrument (blood type) is uncorrelated with any *other* determinants of the outcome (survival)

Suppose that Y denotes the outcome variable, A denotes the treatment variable, and Z denotes the instrumental variable. The effect of the treatment A on the outcome Y that we are interested in is β_1 in the following formulation

$$Y = \beta_0 + \beta_1 A + \epsilon$$

However, we cannot simply use Least Squares (LS) to estimate β_1 because the treatment A is endogenous ($cov(A, \epsilon) \neq 0$) due to unmeasured confounders (measured confounders, on the other hand, can be dealt with by including them in the regression).

To use an instrument to estimate β_1, our instrument must satisfy the relevance criteria:

$$cov(A, Z) \neq 0$$

i.e. the instrument is correlated with the endogenous variable. And it must also satisfy the exclusion restriction:

$$cov(Z, \epsilon) = 0$$

i.e. the instrument is not correlated with the outcome or any other unobserved determinant of it.

If our instrument satisfies these requirements, then we can estimate the treatment effect using two stage least squares (2SLS) [5]. The first stage of 2SLS involves regressing the treatment variable on the instrument

$$A = \alpha_0 + \alpha_1 Z + \gamma$$

The predicted treatment from the first stage LS model, $\hat{A}$, will thus correspond only to the portion of the treatment that is explained to the instrument. The second stage of 2SLS involves regressing the outcome on this predicted treatment:

$$Y = \beta_0 + \beta_1 \hat{A} + \epsilon$$

The upshot is that $\hat{\beta}_1$ will then capture the effect of the parts of the treatment that are influenced by the instrument on the outcome.

Before we're ready to implement 2SLS, we need to first confirm that blood type is a true instrument. So that we don't need to worry about censorship in our outcome (survival time from transplantation), we consider a binary outcome that is survival 3 months from time 0 (MELD 18).

Relevance: blood type is correlated with wait time

Figure 6.6 shows the relative wait times (from MELD 18) in terms of both (a) days and (b) transplant MELD score by recipient blood type. With either metric of wait time, Type AB patients get transplanted faster than all other blood types.

Permutation tests comparing the average wait time in blood groups O and AB, as well as comparing average transplant MELD score in blood groups O and AB each yielded extremely small p-values. Thus it certainly appears that blood type (the instrument) is correlated with wait time (the treatment).

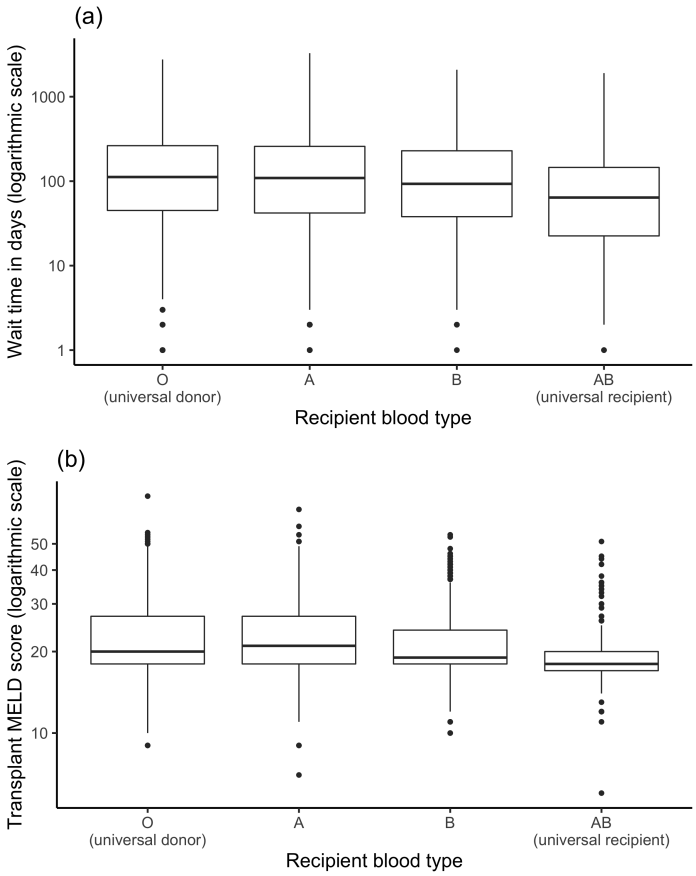

Figure 6.6: Boxplots displaying the distribution by blood type of wait time in terms of (a) days from MELD 18, and (b) transplant MELD score.

Exclusion restriction: Blood type is only related to survival (outcome) only through wait time

There are a range of studies looking at the correlation of blood type with life expectancy (unrelated to transplantation). For instance, there exist studies that imply that blood type B is correlated with higher life expectency [90], that imply that blood type B is correlated with lower life expectency [12], and that there is no correlation between blood type and life expectancy [100]. The lack of agreement between these studies implies a lack of a relationship between blood type and life expectancy. Moreover, blood type is thought to be randomly distributed throughout the population, conditional on family.

There are however some links between blood type and race, and there exist links between race and life expectancy [96, 16]. Figure 6.7 shows the distribution of blood type by race in the data, and is in line with the literature on the topic. Fortunately, we can deal with the correlation of life expectency on race and thus with blood type by conditioning our analysis

on ethnicity (e.g. by including ethnicity as a variable in our regressions).

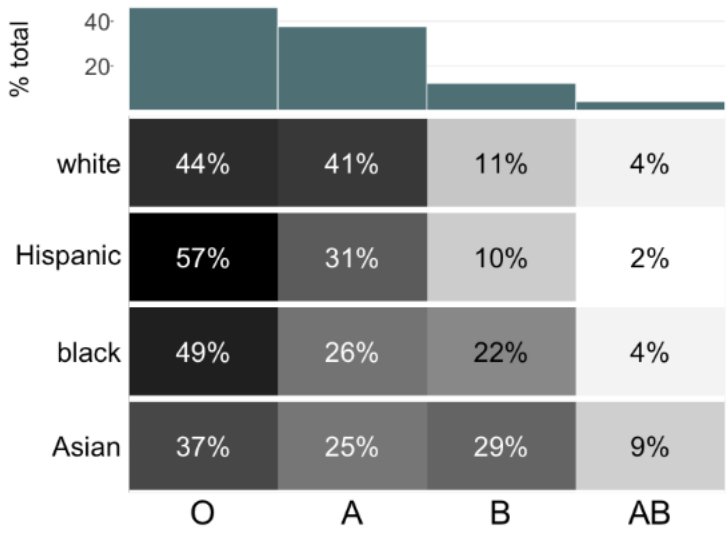

Figure 6.7: A superheatmap displaying the distribution of blood type by race.

6.4 Estimating the effects of wait time on survival using sequential 2SLS

To estimate the effect of wait time on survival, we will use a unique approach from [83] based on 2SLS. This approach, which we call sequential 2SLS, uses a series of two-stage least squares models, one for each additional month of wait time (i.e. we estimate the effect of multiple treatments). We will use two slightly different approaches to estimate two related but different quantities:

1. The effect on death by month t of receiving a transplant by month t versus not having received a transplant by month t. This is known as the failure function
2. The effect on death by month t of being transplanted a month earlier than actually transplanted (given that the transplantation took place some time before month t).

Approach 1: estimating the failure function

The first approach involves estimating the "failure function" at each month t (where time starts at MELD 18), where a "failure" at time t is death by month t. The failure function at time t is the effect on death by month t of receiving a transplant by month t (versus not having received a transplant by month t).

We will consider up to month $t = 24$ where $t = 0$ is the first instance of receiving a MELD score of 18 as discussed above. Suppose that we have a patient who was transplanted at month 3, and died in month 5. Then the treatment vector for this patient is denoted $A = (A_1, ..., A_{24})$, where

$$A_t = 1(\text{transplanted by month } t)$$

Since our patient was transplanted at month 3, their treatment vector is given by $A = (0, 0, 1, 1, ..., 1)$ (i.e. the first two entries are 0 for "untransplanted", and the remainder are 1 for "transplanted").

Their outcome vector is denoted $Y = (Y_1, ..., Y_{24})$, where

$$Y_t = 1(\text{died by month } t)$$

and since our patient died in month 5, their outcome vector is given by $Y = (0, 0, 0, 0, 1, 1, ..., 1)$, i.e. the first four entries are 0 for "alive", and the remainder are 1 for "deceased".

We then estimate 24 models using 2SLS, one for each month up to 24 months. The target for each month is β_t in the formulation below

$$(\text{death by } t \text{ months}) = \alpha_t + \beta_t(\text{transplanted by } t \text{ months}) + \epsilon_t$$

for $t = 1, ..., 24$, and where ϵ_t is an error term. Using our notation, we can write this more compactly as

$$Y_t = \alpha_t + \beta_t A_t + \epsilon_t$$

We estimate β_t for each $t = 1, ..., 24$ using 2SLS with blood type as the instrument. We also adjust by ethnicity and region by including them as variables in both stages of the regression.

The first stage of the 2SLS procedure involves regressing the treatment (A_t, transplantation by month t) on the instrument (blood type dummy variables):

$$A_t = \gamma_{0,t} + \gamma_{1,t}\text{aboA} + \gamma_{2,t}\text{aboB} + \gamma_{3,t}\text{aboAB} + \gamma_{4,t}\text{region} + \gamma_{5,t}\text{ethnicity} + \nu_t$$

Then taking the fitted treatment value from this regression, $\widehat{A}_t$ and using it in the second stage model of the outcome (Y_t, death by month t).

$$Y_t = \beta_{0,t} + \beta_{1,t}\widehat{bloodA}_t + \beta_{2,t}\text{region} + \beta_{3,t}\text{ethnicity} + \epsilon_t$$

The idea is that by first regressing the treatment on the instrument and using the predicted version of the treatment in the outcome model, we are only estimating the effect on the outcome of the portion of the treatment that is influenced by the instrument (which will lead to a valid estimate of the causal effect without any of the confounding).

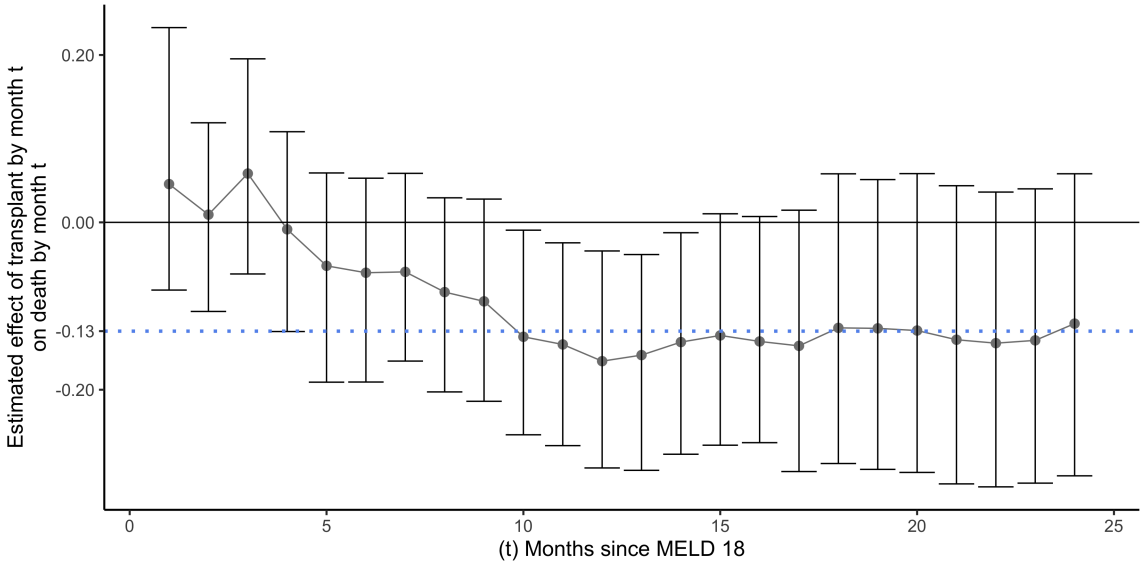

Figure 6.8: The estimated effects of transplantation by month t on death by month t with 95% bootstrapped confidence intervals.

Having fit 24 models, we plot the 24 2SLS causal estimates, $\beta_{1,t}$ for $t = 1, ..., 24$, in Figure 6.8 to get an idea of how the effect changes with time. For instance, the average effect on having died at some point within 24 months if a transplant had been received *at some point* in the 24 month period - versus not having received a transplant yet - is approximately 0.13. This means that patients are on average 13% less likely to have died by 24 months if they received a transplant by 24 months versus if they haven't yet received a transplant. This indicates that there certainly is a substantial benefit in terms of survival of receiving a transplant, which is as expected.

Figure 6.9 shows how these results change if we redefine time 0 by choosing a different MELD score to start time. It is clear that the results are fairly consistent for MELD scores of 16 through 20, but that the effects are less when time 0 is MELD 15.

So while we have shown that there is a survival benefit in terms of survival of having received a transplant versus not having received a transplant yet, we haven't really answered the question of whether there is an observable survival benefit of receiving a transplant *sooner rather than later*. This is what we explore in the next section.

Approach 2

To answer the question of what is the average survival benefit of receiving a transplant one month earlier, we need to change the estimand. Instead of estimating the effect on death by t months of transplantation by t months, this approach estimates the effect on death by t months of being transplanted one month earlier than when the transplant occurred (if it occurred prior to t months).

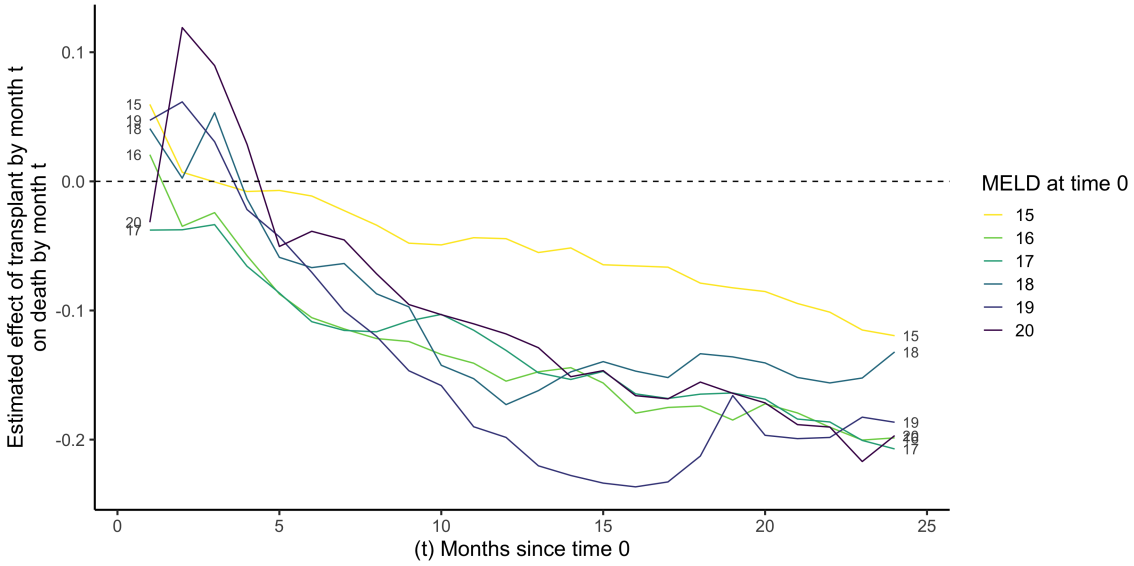

Figure 6.9: The estimated effects of transplantation by month t on death by month t for each definition of time 0, ranging from MELD 15 through to MELD 20. Each line is colored by the time 0 MELD score.

In this approach, the outcome Y_t is defined in the same way as before

$$Y_t = 1(\text{death by month } t)$$

However, this time, the treatment at month t is defined to be

$$A_t = \begin{cases} 0 & \text{if not transplanted by month } t \\ t - \text{transplant month} & \text{if transplanted by month } t \end{cases}$$

So that if we again have a patient who was transplanted at month 3 and who died at month 5, their treatment vector would be

$$A = (A_1, A_2, ..., A_{24}) = (0, 0, 1, 2, 3, 4, ..., 22),$$

and their outcome vector would be

$$Y = (Y_1, Y_2, ..., Y_{24}) = (0, 0, 0, 0, 1, 1, 1, ..., 1).$$

Again we estimate 24 models using 2SLS. However, this time the treatment variable is "months since transplantation" (i.e. t - transplantation month) rather than "transplanted by month t". The models we estimate are as follows:

$$(\text{death by } t \text{ months}) = \alpha_t + \beta_t(\text{months since transplantation}) + \epsilon_t$$

for $t = 1, ..., 24$. Which, using our notation, is equivalent to

$$Y_t = \alpha_t + \beta_t A_t + \epsilon_t$$

The quantity of interest is again β_t for $t = 1, ..., 24$, and this corresponds to the effect of being transplanted one month earlier on death by time t. This is again estimated using 2SLS with blood type as the instrument. We also adjust by region and ethnicity by including them as variables in each stage of the 2SLS regression. In fact, the 2SLS procedure is exactly the same as in the previous section, except with the new treatment variable.

Thus since we again have 24 models across different values of t, we again plot the 2SLS causal estimates, $\beta_{1,t}$ for $t = 1, ..., 24$ to get an idea of how increasing the actual wait time influences the survival in Figure 6.10.

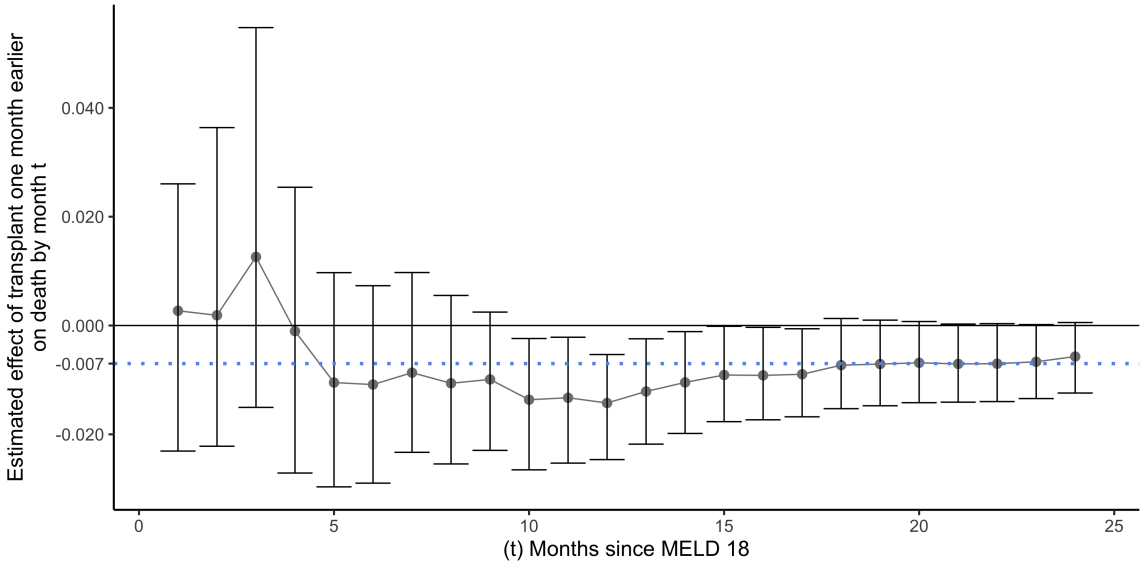

Figure 6.10: The estimated effect of being transplanted one month earlier than transplantation actually occurred on death by month t with 95% bootstrapped confidence intervals.

One way to interpret these results is that the average effect of being transplanted one month earlier is a decrease of 0.7% in the chance of death by 24 months. This can be further interpreted as the average effect of being transplanted 6 months earlier is a decrease of 4.2% ($4.2 = 6 \times 0.7$) in the chance of death by 24 months, which is quite a substantial decrease.

Figure 6.11 dhosws that these results remain relatively consistent across different time 0 MELD score definitions.

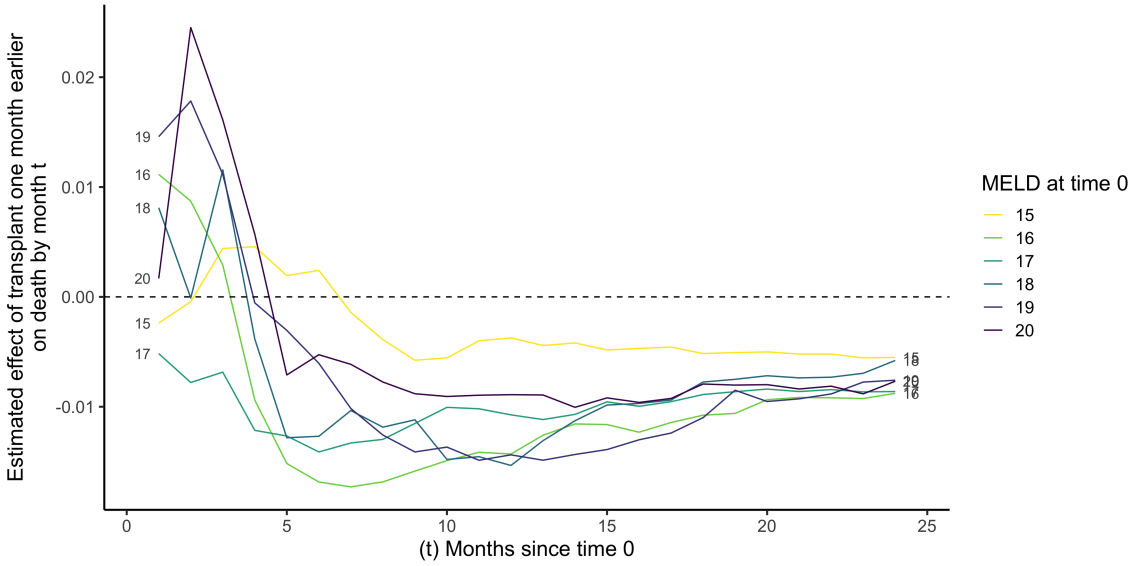

Figure 6.11: The estimated effects of being transplanted one month earlier than transplantation actually occurred on death by month t for each definition of time 0 ranging from MELD 15 through to MELD 20. Each line is colored by the time 0 MELD score, and is also annotated directly with the time 0 MELD score at the start and end point of the line.

6.5 Discussion

This Chapter serves to (1) re-frame of the survival benefit discussion away from comparing transplant outcomes to transplant-free outcomes, in the direction of comparing outcomes under different wait times, and (2) to show that transplant wait time does indeed affect survival.

While this is a fundamentally interesting finding, we are not at the stage where such a result can be used to estimate survival benefit for any individual patient, and thus this approach cannot be used to directly allocate organs.

Moreover, there are no studies as of yet, including our own, that focuses on quality of life, as opposed to years of survival [87].

6.6 Conclusion

We have shown that there is a survival benefit that is conferred by receiving a transplant earlier. For instance, we have shown that if everyone received a transplant 6 months earlier than they actually received a transplant, then the average reduction in death rate by 24 months would be 4.2%.

While we have not developed a method that can be used to estimate the effect of receiving a transplant earlier for any *individual* patient, we have shown that there is an effect of reduced wait-time overall.

Appendix A

Appendix

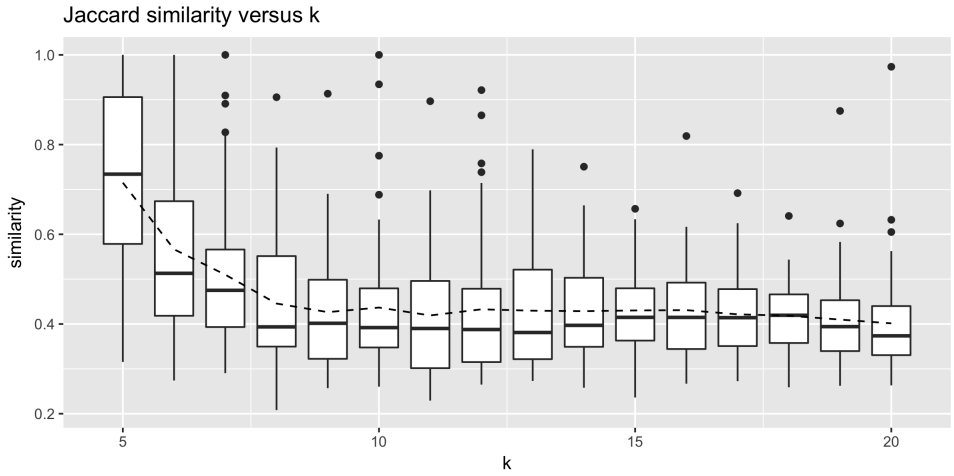

Figure A.1: Average pairwise Jaccard Similarity between 100 90% subsamples of the set of word vectors.

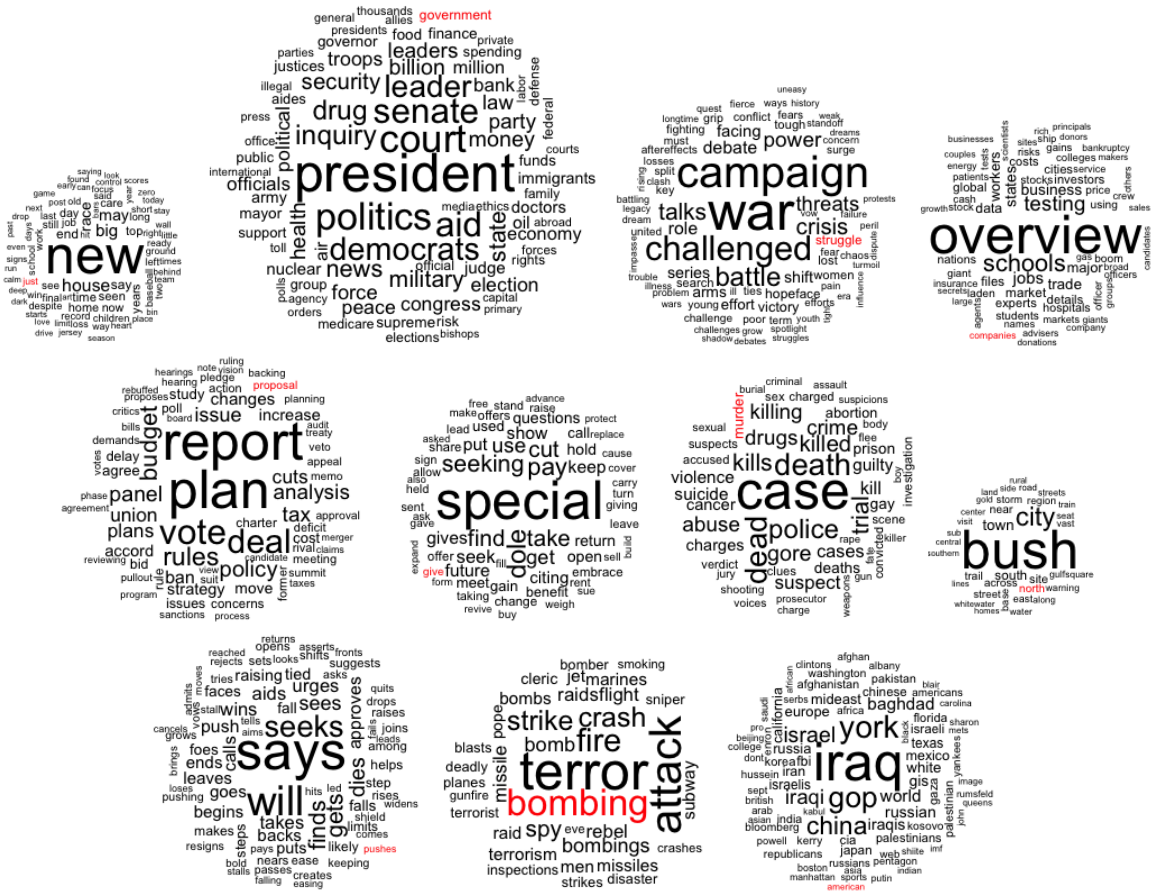

Figure A.2: Word clouds for the 11 word clusters. The word corresponding to the cluster center is highlighted in red. The size of each word corresponds to its frequency in the NY Times headlines corpus.

Figure A.3: Four randomly selected examples of validation images from the top cluster of images in Figure 11.

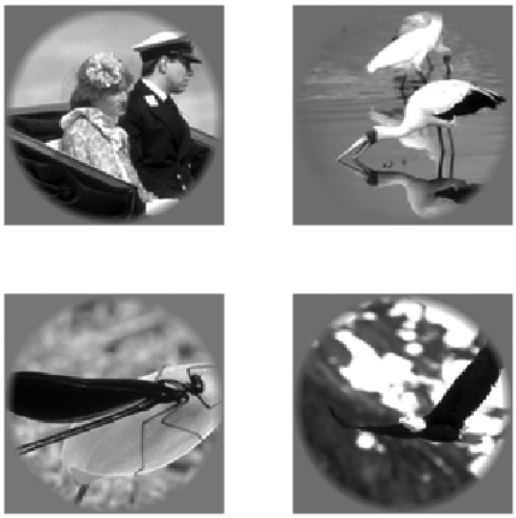

Figure A.4: Four randomly selected examples of validation images from the bottom cluster of images in Figure 11.

PROCCODE	Description	Risk
AAA	Abdominal Aortic Aneurysm	Med Risk
AMP	Limb Amputation	Med Risk
APPY	Appendix	High Risk
AVSD	Av Shunt Dialysis	Low Risk
BILI	Bile Duct Liver Pancreatic	High Risk
BRST	Breast	Low Risk
CARD	Cardiac	Low Risk
CBGB	Coronary Bypass Chest Donor Incision	Low Risk
CBGC	Coronary Bypass Graft Chest Incision	Low Risk
CEA	Carotid Endarterectomy	Low Risk
CHOL	Gallbladder	High Risk
COLO	Colon	High Risk
CRAN	Craniotomy	Low Risk
CSEC	Cesarean Section	Med Risk
FUSN	Spinal Fusion	Low Risk
FX	Fracture	Low Risk
GAST	Gastric	Med Risk
HER	Herniorrhaphy	High Risk
HPRO	Hip Prosthesis	Low Risk
HYST	Abdominal Hysterectomy	Med Risk
KPRO	Knee Prosthesis	Low Risk
KTP	Kidney Transplant	High Risk
LAM	Laminectomy	Low Risk
NECK	Neck	Med Risk
NEPH	Kidney Surgery	Med Risk
OVRY	Ovarian	Med Risk
PACE	Pacemaker	Med Risk
PRST	Prostate	Med Risk
PVBY	Peripheral Vascular Bypass	Low Risk
REC	Rectal	High Risk
RFUSN	Refusion Spine	Low Risk
SB	Small Bowel	High Risk
SPLE	Spleen	Med Risk
THOR	Thoracic	Med Risk
THYR	Thyroid	Low Risk
VHYS	Vaginal Hysterectomy	Med Risk
VSHN	Ventricular Shunt	Low Risk
XLAP	Exploratory Abdominal	High Risk

Table A.1: The list of procedure codes.

Lab	Description
Alanine transferase (ALT)	An enzyme found mostly in liver and kidney cells released into the blood when the liver is damaged
Albumin	A protein made by the liver which can be indicative of liver function
Alkaline phoshatase (ALP)	An enzyme found in several tissues throughout the body (most is found in the bone and liver).
Aspartate transaminase (APT)	An enzyme released when your liver or muscles are damaged
Basophils ABS	A type of white blood cell that fights infection
Bilirubin total	A compound that breaks down heme in vertebrates
Calcium	An element found in the bones, heart, nerves, kidneys, and teeth.
Carbon dioxide	A gas found in the blood that can be indicative of kidney and respiratory problems
Chloride	An element that helps balance the amount of fluid inside and outside of cells
C-Reactive protein (CRP)	A protein made by the liver that is sent into the bloodstream in response to inflammation
Creatinine serum	A waste product in the blood produced by the kidney
E-GFR	Estimated glomerular filtration rate measures the level of kidney function
Glucose	Often used to help diagnose or monitor diabetes
Hematocrit	Red blood cells
Hemoglobin	A protein found in red blood cells that carries oxygen from your lungs to the rest of your body
Lymphocytes ABS	A type of white blood cell
Monocytes ABS	A type of white blood cell
Neutrophil ABS	A type of white blood cell
Platelet count	Tiny fragments of cells that are essential for normal blood clotting
Potassium	An electrolyte essential for proper muscle and nerve function
Protein	A protein test can help diagnose liver and kidney diseases
Red cell count	Red blood cells
Sodium	A mineral particularly important for nerve and muscle function
Urea nitrogen blood (BUN)	The nitrogen in your blood that comes from the waste product urea indicative of kidney function
White cell count	Total number of white blood cells

Table A.2: The list of lab measurements and what they measure.

Elixhauser Category
Rheumatoid Arthritis Collagen Vascular Diseases
Valvular Disease
Liver Disease
Hiv Aids
Solid Tumor Without Metastasis
Metastatic Cancer
Lymphoma
Blood Loss Anemia
Deficiency Anemias
Coagulopathy
Hypothyroidism
Diabetes Uncomplicated
Diabetes Complicated
Peripheral Vascular Disorders
Weight Loss
Obesity
Other Neurological Disorders
Fluid And Electrolye Disorders
Alcohol Abuse
Drug Abuse
Psychoses
Depression
Paralysis
Congestive Heart Failure
Hypertension Combined
Renal Failure
Pulmonary Circulation Disorders
Chronic Pulmonary Disease
Peptic Ulcer Disease Excluding Bleeding

Table A.3: The list of Elixhauser categories.

Medication therapeutic class
Analgesic And Antihistamine Combination
Analgesics
Anesthetics
Anti Obesity Drugs
Antiallergy
Antiarthritics
Antiasthmatics
Antibiotics
Anticoagulants
Antidotes
Antifungals
Antihistamine And Decongestant Combination
Antihistamines
Antihyperglycemics
Antiinfectives
Antiinfectives Miscellaneous
Antiinflam Tumor Necrosis Factor Inhibiting Agents
Antineoplastics
Antiparasitics
Antiparkinson Drugs
Antiplatelet Drugs
Antivirals
Autonomic Drugs
Biologicals
Blood
Cardiac Drugs
Cardiovascular
CNS Drugs
Colony Stimulating Factors
Contraceptives
Cough Cold Preparations
Diagnostic
Diuretics
EENT Preps
Elect Caloric H2O
Gastrointestinal
Herbals
Hormones
Immunosuppressants
Miscellaneous Medical Supplies Devices Non Drug
Muscle Relaxants
Other
Pre Natal Vitamins
Psychotherapeutic Drugs
Sedative Hypnotics
Skin Preps
Smoking Deterrents
Thyroid Preps
Unclassified Drug Products
Vitamins

Table A.4: The list of Medication therapeutic classes.

Bibliography

1. Abouna, G. M. Organ Shortage Crisis: Problems and Possible Solutions. *Transplantation Proceedings* **40,** 34–38. ISSN: 0041-1345 (2008).
2. Aga, E. *et al.* Surgical site infections after abdominal surgery: incidence and risk factors. A prospective cohort study. eng. *Infectious Diseases (London, England)* **47,** 761–767. ISSN: 2374-4243 (2015).
3. Anderson, D. J. *et al.* Strategies to Prevent Surgical Site Infections in Acute Care Hospitals: 2014 Update. *Infection control and hospital epidemiology : the official journal of the Society of Hospital Epidemiologists of America* **35,** 605–627. ISSN: 0899-823X (2014).
4. Andrews, D. F. Plots of High-Dimensional Data. *Biometrics* **28,** 125–136 (1972).
5. Angrist, J. D. & Imbens, G. W. Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity. *Journal of the American Statistical Association* **90,** 431–442. ISSN: 0162-1459 (1995).
6. Angrist, J. D., Imbens, G. W. & Rubin, D. B. Identification of Causal Effects Using Instrumental Variables. *Journal of the American Statistical Association* **91,** 444–455. ISSN: 0162-1459 (1996).
7. Angrist, J. D. & Krueger, A. B. Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments. *Journal of Economic Perspectives* **15,** 69–85. ISSN: 0895-3309 (2001).
8. Barter, R. L. & Yu, B. Superheat: An R Package for Creating Beautiful and Extendable Heatmaps for Visualizing Complex Data. *Journal of Computational and Graphical Statistics* **27,** 910–922. ISSN: 1061-8600 (2018).
9. Basu, S., Kumbier, K., Brown, J. B. & Yu, B. Iterative random forests to discover predictive and stable high-order interactions. *Proceedings of the National Academy of Sciences* **115,** 1943–1948 (2018).
10. Biggins, S. W. & Bambha, K. MELD-based liver allocation: who is underserved? eng. *Seminars in Liver Disease* **26,** 211–220. ISSN: 0272-8087 (2006).
11. Bordeianou, L. *et al.* Truth in Reporting: How Data Capture Methods Obfuscate Actual Surgical Site Infection Rates within a Healthcare Network System. *Diseases of the colon and rectum* **60,** 96–106. ISSN: 0012-3706 (2017).

12. Brecher, M. E. & Hay, S. N. ABO Blood Type and Longevity. en. *American Journal of Clinical Pathology* **135,** 96–98. ISSN: 0002-9173 (2011).

13. Breiman, L. Random Forests. en. *Machine Learning* **45,** 5–32. ISSN: 1573-0565 (2001).

14. Brinton, W. Graphic Methods for Presenting Facts, *New York: The Engineering Magazine Company* (1914).

15. Bujack, R. *et al.* The Good, the Bad, and the Ugly: A Theoretical Framework for the Assessment of Continuous Colormaps. *IEEE Transactions on Visualization and Computer Graphics* **24,** 923–933. ISSN: 1077-2626 (2018).

16. Cantu, P. A., Hayward, M. D., Hummer, R. A. & Chiu, C.-T. New estimates of racial/ethnic differences in life expectancy with chronic morbidity and functional loss: evidence from the National Health Interview Survey. eng. *Journal of Cross-Cultural Gerontology* **28,** 283–297. ISSN: 1573-0719 (2013).

17. Carr, D. B., Littlefield, R. J., Nicholson, W. L. & Littlefield, J. S. Scatterplot Matrix Techniques for Large N. **82,** 424–436 (1987).

18. Chawla, N. V., Japkowicz, N. & Kotcz, A. Editorial: Special Issue on Learning from Imbalanced Data Sets. *SIGKDD Explor. Newsl.* **6,** 1–6. ISSN: 1931-0145 (2004).

19. Chen, C. Generalized Association Plots: Information Visualization via Iteratively Generated Correlation Matrices. *Statistica Sinica,* 7–29 (2002).

20. Chen, C. & Breiman, L. Using Random Forest to Learn Imbalanced Data. *Thechnical Report, University of California, Berkeley* (2004).

21. Cleveland, W. S. *Visualizing Data* (At&T Bell Laboratories, 1993).

22. Cook, D. *et al.* Exploring gene expression data, using plots. *Journal of Data Science* **5,** 151 (2007).

23. Culver, D. H. *et al.* Surgical wound infection rates by wound class, operative procedure, and patient risk index. National Nosocomial Infections Surveillance System. eng. *The American Journal of Medicine* **91,** 152S–157S. ISSN: 0002-9343 (1991).

24. De Lissovoy, G. *et al.* Surgical site infection: incidence and impact on hospital utilization and treatment costs. eng. *American Journal of Infection Control* **37,** 387–397. ISSN: 1527-3296 (2009).

25. Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *PNAS* **95,** 14863–14868 (1998).

26. Englschalk, C. *et al.* Benefit in liver transplantation: a survey among medical staff, patients, medical students and non-medical university staff and students. eng. *BMC medical ethics* **19,** 7. ISSN: 1472-6939 (2018).

27. Fayek, S. A., Quintini, C., Chavin, K. D. & Marsh, C. L. The Current State of Liver Transplantation in the United States. en. *American Journal of Transplantation* **16,** 3093–3104. ISSN: 1600-6143 (2016).

28. Frangakis, C. E. & Rubin, D. B. Principal Stratification in Causal Inference. *Biometrics* **58,** 21–29. ISSN: 0006-341X (2002).
29. Galili, T., O'Callaghan, A., Sidi, J. & Sievert, C. heatmaply: an R package for creating interactive cluster heatmaps for online publishing. en. *Bioinformatics* (2017).
30. Garcia, G. G., Harden, P. N. & Chapman, J. R. The global role of kidney transplantation. English. *Kidney International* **81,** 425–427. ISSN: 0085-2538 (2012).
31. Gaujoux, R. & Seoighe, C. A flexible R package for nonnegative matrix factorization. eng. *BMC bioinformatics* **11,** 367. ISSN: 1471-2105 (2010).
32. Gianfrancesco, M. A., Tamang, S., Yazdany, J. & Schmajuk, G. Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data. *JAMA internal medicine* **178,** 1544–1547. ISSN: 2168-6106 (2018).
33. GODT. *Global Observatory on Donation and Transplantation Database* 2016.
34. Google. *Google Code Archive* 2013. <`https://code.google.com/archive/p/word2vec/`> (visited on 02/20/2018).
35. Gower, J. & Digby, P. Expressing Complex Relationships in Two Dimensions. *Interpreting Multivariate Data, ed. V. Barnett, Chichester, U.K.: Wiley,* 83–118 (1981).
36. Greenland, S., Robins, J. M. & Pearl, J. Confounding and Collapsibility in Causal Inference. en. *Statistical Science* **14,** 29–46. ISSN: 0883-4237, 2168-8745 (1999).
37. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. en. *Bioinformatics* **32,** 2847–2849. ISSN: 1367-4803 (2016).
38. Haley, R. W. *et al.* Identifying patients at high risk of surgical wound infection. A simple multivariate index of patient susceptibility and wound contamination. eng. *American Journal of Epidemiology* **121,** 206–215. ISSN: 0002-9262 (1985).
39. Harrower, M. & Brewer, C. A. ColorBrewer.org: An Online Tool for Selecting Colour Schemes for Maps. *The Cartographic Journal* **40,** 27–37. ISSN: 0008-7041 (2003).
40. Hothorn, T., Hornik, K., Wiel, M. A. v. d. & Zeileis, A. A Lego System for Conditional Inference. *The American Statistician* **60,** 257–263. ISSN: 0003-1305 (2006).
41. Inselberg, A. en. in *Encyclopedia of Database Systems* (eds LIU, L. & ÖZSU, M. T.) 2018 (Springer US, Boston, MA, 2009). ISBN: 978-0-387-39940-9. doi:`10.1007/978-0-387-39940-9_262`.
42. Inselberg, A. The plane with parallel coordinates. *The Visual Computer* **1,** 69–91 (1985).
43. Inselberg, A. *Visual Data Mining with Parallel Coordinates* SSRN Scholarly Paper ID 85868 (Social Science Research Network, Rochester, NY, 1998).
44. Inselberg, A. & Dimsdale, B. in *Computer Graphics 1987* (ed Kunii, D. T. L.) 25–44 (Springer Japan, 1987).

45. Jacob, M. *et al.* Pretransplant MELD score and post liver transplantation survival in the UK and Ireland. en. *Liver Transplantation* **10,** 903–907. ISSN: 1527-6473 (2004).

46. Kamath, P. S. *et al.* A model to predict survival in patients with end-stage liver disease. eng. *Hepatology (Baltimore, Md.)* **33,** 464–470. ISSN: 0270-9139 (2001).

47. Kaufman, L. & Rousseeuw, P. J. en. in *Finding Groups in Data* 68–125 (John Wiley & Sons, Inc., 1990). ISBN: 978-0-470-31680-1.

48. Kay, K. N., Naselaris, T., Prenger, R. J. & Gallant, J. L. Identifying natural images from human brain activity. *Nature* **452,** 352–355. ISSN: 0028-0836 (2008).

49. Ke, C. *et al.* Prognostics of surgical site infections using dynamic health data. eng. *Journal of Biomedical Informatics* **65,** 22–33. ISSN: 1532-0480 (2017).

50. Keller, E. J., Kwo, P. Y. & Helft, P. R. Ethical considerations surrounding survival benefit–based liver allocation. en. *Liver Transplantation* **20,** 140–146. ISSN: 1527-6473 (2014).

51. Kim, S.-Y. Effects of sample size on robustness and prediction accuracy of a prognostic gene signature. eng. *BMC bioinformatics* **10,** 147. ISSN: 1471-2105 (2009).

52. Kolde, R. Pheatmap: pretty heatmaps. *R package version* **61** (2012).

53. Korol, E. *et al.* A Systematic Review of Risk Factors Associated with Surgical Site Infections among Surgical Patients. *PLoS ONE* **8.** ISSN: 1932-6203 (2013).

54. Kremers, W. K. *et al.* MELD score as a predictor of pretransplant and posttransplant survival in OPTN/UNOS status 1 patients. en. *Hepatology* **39,** 764–769. ISSN: 1527-3350 (2004).

55. Kumbier, K., Basu, S., Brown, J. B., Celniker, S. & Yu, B. Refining interaction search through signed iterative Random Forests. *arXiv:1810.07287 [cs, stat].* arXiv: 1810.07287. (Visited on 10/28/2019) (2018).

56. Lai, Q. *et al.* Intention-to-treat survival benefit of liver transplantation in patients with hepatocellular cancer. eng. *Hepatology (Baltimore, Md.)* **66,** 1910–1919. ISSN: 1527-3350 (2017).

57. Lawson, E. H., Hall, B. L. & Ko, C. Y. Risk factors for superficial vs deep/organ-space surgical site infections: implications for quality improvement initiatives. eng. *JAMA surgery* **148,** 849–858. ISSN: 2168-6262 (2013).

58. Lee, T. S. Image representation using 2D Gabor wavelets. *IEE Transactions on Pattern Analysis and Machine Intelligence* **18,** 959–971 (1996).

59. Ling, R. A Computer Generated Aid for Cluster Analysis. *Communications of the ACM,* 355–361 (1973).

60. Liu, X.-Y., Wu, J. & Zhou, Z.-H. Exploratory Undersampling for Class-Imbalance Learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* **39,** 539–550. ISSN: 1083-4419, 1941-0492 (2009).

61. London, A. J. Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability. en. *Hastings Center Report* **49,** 15–21. ISSN: 1552-146X (2019).

62. Loua, T. Atlas statistique de la population de Paris (1873).

63. Maaten, L. v. d. & Hinton, G. Visualizing Data using t-SNE. *Journal of Machine Learning Research* **9,** 2579–2605. ISSN: ISSN 1533-7928 (2008).

64. Magill, S. S. *et al.* Prevalence of healthcare-associated infections in acute care hospitals in Jacksonville, Florida. eng. *Infection Control and Hospital Epidemiology* **33,** 283–291. ISSN: 1559-6834 (2012).

65. Malinchoc, M. *et al.* A model to predict poor survival in patients undergoing transjugular intrahepatic portosystemic shunts. eng. *Hepatology (Baltimore, Md.)* **31,** 864–871. ISSN: 0270-9139 (2000).

66. Manrai, A. K. *et al.* Genetic Misdiagnoses and the Potential for Health Disparities. *New England Journal of Medicine* **375,** 655–665. ISSN: 0028-4793 (2016).

67. Martin, A. P., Bartels, M., Hauss, J. & Fangmann, J. Overview of the MELD Score and the UNOS Adult Liver Allocation System. *Transplantation Proceedings* **39,** 3169–3174. ISSN: 0041-1345 (2007).

68. Merion, R. M. *et al.* The survival benefit of liver transplantation. eng. *American Journal of Transplantation: Official Journal of the American Society of Transplantation and the American Society of Transplant Surgeons* **5,** 307–313. ISSN: 1600-6135 (2005).

69. Mikolov, T., Chen, K., Corrado, G. & Dean, J. Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781.* arXiv: 1301.3781 (2013).

70. Moylan, C. A. *et al.* Disparities in Liver Transplantation Before and After Introduction of the MELD Score. en. *JAMA* **300,** 2371–2378. ISSN: 0098-7484 (2008).

71. Mu, Y., Edwards, J. R., Horan, T. C., Berrios-Torres, S. I. & Fridkin, S. K. Improving Risk-Adjusted Measures of Surgical Site Infection for the National Healthcare Safely Network. en. *Infection Control & Hospital Epidemiology* **32,** 970–986. ISSN: 0899-823X, 1559-6834 (2011).

72. Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R. & Yu, B. Interpretable machine learning: definitions, methods, and applications. *arXiv:1901.04592.* arXiv: 1901.04592 (2019).

73. Neuberger, J. An update on liver transplantation: A critical review. eng. *Journal of Autoimmunity* **66,** 51–59. ISSN: 1095-9157 (2016).

74. Nussbaumer Knaflic, C. *Storytelling with Data: A Data Visualization Guide for Business Professionals* 1 edition. English. ISBN: 978-1-119-00225-3 (Wiley, Hoboken, New Jersey, 2015).

75. O'Dell, H. W. *et al.* Public attitudes toward contemporary issues in liver allocation. eng. *American Journal of Transplantation: Official Journal of the American Society of Transplantation and the American Society of Transplant Surgeons* **19,** 1212–1217. ISSN: 1600-6143 (2019).

76. Of Medicine of the National Academies, I. *Organ Donation: Opportunities for Action* en. ISBN: 978-0-309-10114-1. doi:10.17226/11643 (2006).

77. Pan, S. J. & Yang, Q. A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering* **22,** 1345–1359. ISSN: 1041-4347, 1558-2191, 2326-3865 (2010).

78. Papernot, N. *et al. Practical Black-Box Attacks Against Machine Learning* in *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security* event-place: Abu Dhabi, United Arab Emirates (ACM, New York, NY, USA, 2017), 506–519. ISBN: 978-1-4503-4944-4.

79. Price, C. S. & Savitz, L. A. Improving the measurement of surgical site infection risk stratification/outcome detection. *AHRQ* **12** (2012).

80. Provost, F. *Machine Learning from Imbalanced Data Sets 101 (Extended Abstract)* (2000).

81. Rana, A. *et al.* Survival Benefit of Solid-Organ Transplant in the United States. en. *JAMA Surgery* **150,** 252–259. ISSN: 2168-6254 (2015).

82. Reynolds, A. P., Richards, G., Iglesia, B. d. l. & Rayward-Smith, V. J. Clustering Rules: A Comparison of Partitioning and Hierarchical Clustering Algorithms. en. *Journal of Mathematical Modelling and Algorithms* **5,** 475–504. ISSN: 1570-1166, 1572-9214 (2006).

83. Rose, E. & Shem-Tov, Y. *Does Incarceration Increase Crime?* en. SSRN Scholarly Paper ID 3205613 (Social Science Research Network, Rochester, NY, 2018).

84. Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* **20,** 53–65 (1987).

85. Rubin, D. B. Matching to Remove Bias in Observational Studies. *Biometrics* **29,** 159–183. ISSN: 0006-341X (1973).

86. Ruf, A. E. *et al.* Addition of serum sodium into the MELD score predicts waiting list mortality better than MELD alone. en. *Liver Transplantation* **11,** 336–343. ISSN: 1527-6473 (2005).

87. Saab, S. *et al.* MELD fails to measure quality of life in liver transplant candidates. eng. *Liver Transplantation: Official Publication of the American Association for the Study of Liver Diseases and the International Liver Transplantation Society* **11,** 218–223. ISSN: 1527-6465 (2005).

88. Schaubel, D. E. *et al.* Survival Benefit-Based Deceased-Donor Liver Allocation. *American journal of transplantation : official journal of the American Society of Transplantation and the American Society of Transplant Surgeons* **9,** 970–981. ISSN: 1600-6135 (2009).

89. Schep, A. N. & Kummerfeld, S. K. iheatmapr: Interactive complex heatmaps in R. en. *The Journal of Open Source Software* (2017).

90. Shimizu, K. *et al.* Blood type B might imply longevity. *Experimental Gerontology* **39,** 1563–1565. ISSN: 0531-5565 (2004).

91. Soguero-Ruiz, C. *et al.* Data-driven Temporal Prediction of Surgical Site Infection. *AMIA Annual Symposium Proceedings* **2015,** 1164–1173. ISSN: 1942-597X (2015).

92. Stekhoven, D. J. & Bühlmann, P. MissForest—non-parametric missing value imputation for mixed-type data. en. *Bioinformatics* **28,** 112–118. ISSN: 1367-4803 (2012).

93. Steyerberg, E. W. *et al.* Prognosis Research Strategy (PROGRESS) 3: Prognostic Model Research. en. *PLOS Medicine* **10,** e1001381. ISSN: 1549-1676 (2013).

94. Tajima, J. Uniform color scale applications to computer graphics. *Computer Vision, Graphics, and Image Processing* **21,** 305–325. ISSN: 0734-189X (1983).

95. Tardu, M., Bulut, S. & Kavakli, I. H. MerR and ChrR mediate blue light induced photo-oxidative stress response at the transcriptional level in *Vibrio cholerae*. en. *Scientific Reports* **7,** 40817. ISSN: 2045-2322 (2017).

96. Thielke, S. M. *et al.* Sex, Race, and Age Differences in Observed Years of Life, Healthy Life, and Able Life among Older Adults in The Cardiovascular Health Study. *Journal of Personalized Medicine* **5,** 440–451. ISSN: 2075-4426 (2015).

97. Trakhtenberg, E. F. *et al.* Cell types differ in global coordination of splicing and proportion of highly expressed genes. en. *Scientific Reports* **6,** 32249. ISSN: 2045-2322 (2016).

98. Trieu, J. A., Bilal, M. & Hmoud, B. Factors associated with waiting time on the liver transplant list: an analysis of the United Network for Organ Sharing (UNOS) database. *Annals of Gastroenterology* **31,** 84–89. ISSN: 1108-7471 (2018).

99. UNDP. *Human Development Data* 2015.

100. Vasto, S. *et al.* Blood group does not appear to affect longevity a pilot study in centenarians from Western Sicily. en. *Biogerontology* **12,** 467. ISSN: 1389-5729, 1573-6768 (2011).

101. Vock, D. M. *et al.* Assessing the Causal Effect of Organ Transplantation on the Distribution of Residual Lifetime. *Biometrics* **69.** ISSN: 0006-341X (2013).

102. Vu, V. Q. *et al.* Encoding and decoding V1 fMRI responses to natural images with sparse nonparametric models. *Annals of Applied Statistics* **5,** 1159–1182 (2011).

103. Vu, V. Q. *et al.* in *Advances in Neural Information Processing Systems 21* (eds Koller, D., Schuurmans, D., Bengio, Y. & Bottou, L.) 1337–1344 (Curran Associates, Inc., 2009).

104. Walraven, C. v. & Musselman, R. The Surgical Site Infection Risk Score (SSIRS): A Model to Predict the Risk of Surgical Site Infections. en. *PLOS ONE* **8,** e67167. ISSN: 1932-6203 (2013).

105. Wickham, H. A Layered Grammar of Graphics. *Journal of Computational and Graphical Statistics* **19** (2010).

106. Wickham, H. *ggplot2: elegant graphics for data analysis* (Springer, 2016).

107. Wiesner, R. H. *et al.* MELD and PELD: application of survival models to liver allocation. eng. *Liver Transplantation: Official Publication of the American Association for the Study of Liver Diseases and the International Liver Transplantation Society* **7,** 567–580. ISSN: 1527-6465 (2001).

108. Wiesner, R. *et al.* Model for end-stage liver disease (MELD) and allocation of donor livers. eng. *Gastroenterology* **124,** 91–96. ISSN: 0016-5085 (2003).

109. Wilkinson, L. SYSTAT for DOS: Advanced Applications, Version 6. *Evanston, IL: SYSTAT Inc.* (1994).

110. Wilkinson, L. & Friendly, M. The History of the Cluster Heat Map. *The American Statistician* **63,** 179–184 (2009).

111. Wong, B. *Points of view: Avoiding color* en. News. 2011. doi:10.1038/nmeth.1642.

112. Wright, M. N. & Ziegler, A. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. en. *Journal of Statistical Software* **77,** 1–17. ISSN: 1548-7660 (2017).

113. Yeh, H., Smoot, E., Schoenfeld, D. A. & Markmann, J. F. Geographic Inequity in Access to Livers for Transplantation. *Transplantation* **91,** 479–486. ISSN: 0041-1337 (2011).

114. Yu, B. Stability. EN. *Bernoulli* **19,** 1484–1500. ISSN: 1350-7265 (2013).

115. Yu, B. & Kumbier, K. Three principles of data science: predictability, computability, and stability (PCS). *arXiv:1901.08152 [cs, stat].* arXiv: 1901.08152 (2019).